UNITED STATES DISTRICT COURT
WESTERN DISTRICT OF MICHIGAN
SOUTHERN DIVISION

UNITED STATES OF AMERICA,

Plaintiff,

v.

DANIEL GISSANTANER,

Defendant.
/

Case no. 1:17-cr-130

Hon. Janet T. Neff
United States District Judge

Hon. Ray Kent
United States Magistrate Judge

DEFENDANT'S REPLY TO THE GOVERNMENT'S RESPONSE TO MOTION TO EXCLUDE DNA EVIDENCE

NOW COMES, the defendant, Daniel Gissantaner, by and through his attorney, Joanna C. Kloet, Assistant Federal Public Defender, and hereby requests to file this Reply to the Government's Response to the Defendant's Motion to Exclude DNA Evidence, as authorized by Federal Rule of Criminal Procedure 12 and Local Criminal Rules 47.1 and 47.2.

This case was filed in Federal Court after the defendant was acquitted of the charge of felon in possession of a firearm after a July 8, 2016, hearing before the Michigan Department of Corrections, where the only apparent difference between the evidence available to the respective authorities is the DNA likelihood ratio ("LR") now proffered by the Federal Government. Accordingly, the defense submitted the underlying Motion to Exclude DNA Evidence as a dispositive pre-trial motion. Local Criminal Rule 47.1 allows the Court to permit or require further briefing on dispositive motions following a response. Alternatively, if the Court determines the underlying Motion is non-dispositive, the defense requests leave to file this Reply under Local

Criminal Rule 47.2. Good cause exists because with this Reply, the defense seeks to narrow the varied and complex issues that have been raised before the Court, in preparation for the hearing on March 22, 2018.

I. The Likelihood Ratio ("LR") is unreliable and should be excluded under FRE 702 and *Daubert* because the validation studies are insufficient.

The validation studies to which the Government refers in its Response do not demonstrate the STRmix program has been adequately validated and that it is appropriate for use in situations such as the instant matter.[1] To establish foundational validity, "the procedures that comprise [a methodology] must be shown, based on empirical studies, to be repeatable, reproducible, and accurate, at levels that have been measured and are appropriate to the intended application."[2] Validation studies must involve a sufficiently large number of examiners and be based on sufficiently large collections of known and representative samples from relevant populations to reflect the range of features or combinations of features that will occur in the application.[3] Furthermore, the studies should be conducted or overseen by individuals or organizations that have no stake in the outcome of the studies.[4]

In its Response, the Government attaches two studies in support of its contention that STRmix was validated properly.[5] However, these studies were authored by the creator of the STRmix software itself, John S. Buckleton (who also happens to be the witness the Government

[1] *See* ECF No. 52, pp 12-15.
[2] Exhibit 1, additional excerpts from PCAST Forensic Science Report ("PCAST Report Excerpts"), p 47, citing National Physical Laboratory, "A Beginner's Guide to Measurement," and Pavese, F., "An Introduction to Data Modeling Principles in Metrology and Testing," in Data Modeling for Metrology and Testing in Measurement Science," Pavese, F. and A.B. Forbes (Eds.), Birkhauser (2009).
[3] Exhibit 1, PCAST Report Excerpt, p 52.
[4] Exhibit 1, PCAST Report Excerpt, p 52.
[5] ECF No. 52, pp 12-13.

indicated it would seek to call in its favor at the March 22, 2018, hearing before this Court).[6] The Government also cites Buckleton's website for the assertion that STRmix was tested and reliable as shown by 19 publications from 2013 to 2017, but again, Buckleton authored 18 out of 19 of these studies.[7] Moreover, the Government has not shown that these studies were in fact validation studies that closely followed the FBI's SWGDAM Guidelines for Validation of Probabilistic Genotyping Systems – in fact, apparently at least 14 of these studies ***preceded*** the publication of those Guidelines.[8] Likewise, the link the Government provides to Buckleton's personal blog contains links to studies performed by seven local law enforcement agencies, at least three of which were before the FBI SWGDAM Guidelines were issued, and all of which lack evidence of rigorous peer-review.[9] Furthermore, information available on Buckleton's blog indicates that the STRmix software version that was ostensibly "validated" by these local agencies was later subject to numerous revisions and updates to correct errors in the software.[10] Notably, the FBI SWGDAM Guidelines require re-validation following changes to the software that "may impact interpretation or the analytical process."[11]

Even if the existence of validation studies on outdated versions of STRmix provide validation in some factual circumstances, this does not signify that the software is fit for the purpose for which it was employed here.[12] Because "crime laboratories are being asked to evaluate

[6] *See* ECF No. 52-3 and ECF No. 52-4.
[7] ECF No. 52, p 12.
[8] ECF No. 52, p 12; Exhibit 2, FBI SWGDAM Guidelines for Validation of Probabilistic Genotyping Systems, June 15, 2015, p 11.
[9] ECF No. 52, p 12; Exhibit 2, FBI SWGDAM Guidelines for Validation of Probabilistic Genotyping Systems, June 15, 2015, p 11.
[10] Exhibit 3, "A summary of the seven identified miscodes in STRmix," located at https://johnbuckleton.files.wordpress.com/2017/12/a-summary-of-the-seven-identified-miscodes-in-strmix.pdf (last accessed Feb. 23, 2018).
[11] Exhibit 2, FBI SWGDAM Guidelines for Validation of Probabilistic Genotyping Systems, June 15, 2015, p 11.
[12] Exhibit 1, PCAST Report Excerpt, p 56.

many more poor-quality, low-template, and complex DNA mixtures," DNA mixture software "is being used on the most dangerous, least information-rich samples you encounter."[13] "Common characteristics of forensic casework samples that can increase their complexity include multiple contributors, low quantity (provoking possible drop-out) and low quality (e.g., degradation, inhibition, contamination)."[14] "As the number of potential contributors increases, so does uncertainty in accurately determining the true number of contributors."[15] Accuracy may degrade as a function of the absolute and relative amounts of DNA from various contributors.[16] More robust validation is important because it "may determine that, past a certain number of contributors, the information content of the profile is simply too limited to reliably distinguish a true contributor from a non-contributor who shares some of the detected alleles by chance."[17]

The addendum shows that the authors of the PCAST Report responded to its critics, but did not change position.[18] Nor has the report been withdrawn. In preparing the addendum, PCAST reviewed hundreds of papers cited by the respondents and invited the submission of additional studies not considered by PCAST that purport to establish validity and reliability.[19] After undertaking additional study, including convening a meeting with STRmix's creator John

[13] Exhibit 4, Frederick R. Bieber, et al, "Evaluation of forensic DNA mixture evidence: protocol for evaluation, interpretation, and statistical calculations using the combined probability of inclusion," BMC Genetics (2016) 17:125; Joe Palazzolo, "Defense Attorneys Demand Closer Look at Software Used to Detect Crime-Scene DNA," WALL ST. J., Nov. 18, 2015.

[14] Exhibit 5, Hinda Haned, et al., "Validation of probabilistic genotyping software for use in forensic DNA casework: definitions and illustrations," 56 Science and Justice 104, 106 (2016). On a related point, unlike here, the 2015 Michigan state court decision by Muskegon County Circuit Court Judge William C. Marietti in *People v Muhammad* (Case No. 14-65263-FC), involved only an apparent two-person mixture.

[15] Exhibit 4, Frederick R. Bieber, et al, "Evaluation of forensic DNA mixture evidence: protocol for evaluation, interpretation, and statistical calculations using the combined probability of inclusion," BMC Genetics (2016) 17:125.

[16] Exhibit 1, PCAST Report Excerpt, p 79.

[17] Exhibit 5, Hinda Haned, et al., "Validation of probabilistic genotyping software for use in forensic DNA casework: definitions and illustrations," 56 Science and Justice 104, 106 (2016).

[18] Exhibit 6, An Addendum to the PCAST Report on Forensic Science in Criminal Courts, (January 6, 2017) ("PCAST Report Addendum"), pp 1, 8.

[19] Exhibit 6, PCAST Report Addendum, pp 2-3, 5, 8.

Buckleton, PCAST observed that "empirical testing of [probabilistic genotyping systems] had largely been limited to a narrow range of parameters (number and ratios of contributors)," and recommended further testing of "a diverse collection of samples within well-defined ranges."[20] The addendum also stated:

> PCAST has great respect for the value of examiners' experience and judgment: they are critical factors in ensuring that a scientifically valid and reliable method is practiced correctly. However, experience and judgment alone – no matter how great – can *never* establish the validity or degree of reliability of any particular method. Only empirical testing can do so.[21]

In this case, determining the validity of the method requires rigorous software testing and scrutiny of the assumptions underlying the algorithm. Thus, and especially in light of the ongoing changes and updates to the software, validity testing should involve not just forensic scientists and mathematicians, but also software engineers with experience in verification and validation of software. However, the defense is not aware of any fully independent development validation studies conducted on STRmix involving software engineers.[22]

[20] Exhibit 6, PCAST Report Addendum, pp 2-3, 5, 8, 9

[21] Exhibit 6, PCAST Report Addendum, p 3.

[22] Exhibit 5, Hinda Haned, et al, "Validation of probabilistic genotyping software for use in forensic DNA casework: Definitions and illustrations," Science & Justice, 56 (2016) 104-108, pp 106, 107 (Software validation should include "[v]erification of the code itself through visual inspection and recoding," which is "most easily achievable through open-source software.") The apparent lack of thorough validation studies may be attributable to the fact that STRmix's software code is proprietary, a notion inherently incompatible with evidentiary standards. *See* Exhibit 7, Erin E. Kenneally, "Gatekeeping Out of the Box: Open Source Software as a Mechanism to Assess Reliability for Digital Evidence," 6 VA. J.L. & TECH. 13, 149 (2001) ("[c]ross-examining the expert whose opinion is derived from [cases involving software] may call for a cross-examination of the software, which is accurately done by having the source code").

Additionally, the Government's claim that the defense failed to request a re-run of the data is an improper attempt to shift the burden to the defense to prove his own innocence. To establish admissibility for the results generated by the software, the Government is required to demonstrate by a preponderance of the evidence that the evidentiary standards are met. In fact, the accusation that the defendant has not made this request because results would be unfavorable to him is speculative and is just as applicable to the Government – i.e., the Government may not have

While analysis of single-source or simple mixtures of two individuals' DNA is largely an objective method,[23] a great deal of subjectivity is involved in the forensic analysis of complex mixtures of touch DNA – and "[u]sing software doesn't solve the problem, because human biases, assumptions, and discretions go into the software."[24] Even the creators of STRmix did not seriously dispute that "there is no consensus within the forensic biology community as to how [complex mixtures and small DNA samples] should be interpreted."[25] Some complex low-level DNA mixtures are "difficult, or impossible, to interpret reliably."[26]

The FBI concedes that probabilistic genotyping is "not intended to replace the human evaluation of the forensic DNA typing results or the human review of the output prior to reporting," which include the analyst's responsibility to "estimate and use a specific number of contributors in a statistical calculation when interpreting a DNA mixture, or to assess whether typing results should be interpreted or not based on quality."[27] But renowned DNA experts have recognized that in some circumstances, this as a version of the age-old computing truth: "GIGO," or "garbage in, garbage out."[28] The evidentiary standards embedded in our criminal justice system reflect the important proposition that good science requires consensus-based, sound scientific principles, fit for the purpose applied. *See Daubert v. Merrell-Dow Pharm. Inc.*, 509 U.S. 579 (1993); *see also*

re-run the program because the resulting LR would not be favorable to its case. In any case, put simply, the Government, not the defendant, carries the burden.

[23] *See* Exhibit 1, PCAST Report Excerpts, pp 69-83.
[24] Exhibit 8, Lael Henterly, "The Troubling Trial of Emanuel Fair," Seattle Weekly, Jan. 11, 2017.
[25] Exhibit 9, Kelly, H., et al., "A comparison of statistical models for the analysis of complex forensic DNA profiles, Journal of Science and Justice," 54 (2014) (quoting abstract). *See also* Exhibit 10, Seth Augustine, "DNA Mixtures Topic of ISHI Talks, NIST Testing—and Conflict of Interest Accusations," Forensic Magazine, Oct. 5, 2017.
[26] Exhibit 11, "NIST To Assess the Reliability of Forensic Methods for Analyzing DNA Mixtures," Oct. 3, 2017, located at https://www.nist.gov/news-events/news/2017/10/nist-assess-reliability-forensic-methods-analyzing-dna-mixtures (last accessed Feb. 23, 2018).
[27] Exhibit 2, FBI SWGDAM Guidelines for Validation of Probabilistic Genotyping Systems, pp 2, 4.
[28] Exhibit 5, Haned, H., et al, "Validation of probabilistic genotyping software for use in forensic DNA casework: Definitions and illustrations," Science & Justice, 56 (2016) 104-108, p 106.

FRE 702, 703. Being "cautiously optimistic" about the promises that a new science may deliver does not mean the science is ready to be used in every single factual scenario.[29] As PCAST cautioned, additional empirical study of probabilistic genotyping as it applies to complex mixtures is required.[30] For those reasons, at the present time, the PCAST report cautions against its use in cases exactly like the one before the bar.

II. The DNA evidence is more prejudicial than probative and should be excluded pursuant to FRE 403.

This Court's gatekeeping duty under FRE 611(a) sets forth an obligation to ensure the evidence is presented in an effective and efficient manner, such that even relevant evidence can be prohibited under FRE 403 if its prejudicial effect outweighs its probative value. The basic question before the Court is whether it is possible to convey results of forensic DNA analysis in a way that is accurate, reliable, and most importantly, that jurors can understand without the risk of being unfairly prejudiced by it.

As the Supreme Court noted, "[e]xpert evidence can be both powerful and quite misleading because of the difficulty in evaluating it." *Daubert*, *supra*, 509 U.S. at 595. Accordingly, in FRE 403 analyses, a judge "exercises more control over experts than over lay witnesses." *Id*. "One of the major concerns that has been raised about population proportions and statistics on the probability of a match is that jurors will mistakenly assume these statistics directly measure the probability of the defendant's innocence."[31] This phenomenon, known as the Prosecutor's Fallacy,

[29] ECF No. 52, p 22. Relatedly, the Government's accusation that the defense used the *Bonds* case as a "straw man" is inaccurate. *See* ECF No. 52, pp 10-11, n 4. *Bonds* is one of the leading cases in the Sixth Circuit on the admissibility of DNA under the *Daubert* standard. The defense gave a detailed explanation of the technology and the type of biological material at issue in *Bonds* to distinguish it from the matter before the Court in the instant case.

[30] Exhibit 6, PCAST Addendum, p 8.

[31] Exhibit 12, William C. Thompson, "Are Juries Competent to Evaluate Statistical Evidence?" LAW & CONTEMPORARY PROBS, Vol 52: No 4 (1989), pp 25-26.

can and does result in mistaken conclusions.[32] As Professor Laurence Tribe cautioned, "the very mystery that surrounds mathematical arguments – the relative obscurity that makes them at once impenetrable by the layman and impressive to him – creates a continuing risk that he will give such arguments a credence they may not deserve and a weight they cannot logically claim."[33]

The capability of the jurors – not statisticians and forensic examiners – should determine how and if such information is communicated. Generally speaking, a LR involves two factual hypotheses, both of which are created by the prosecution, not the defense, to produce a quantitative statement of probability, which the laboratory then tries to translate into a qualitative statement of probability. The calculation itself involves the application of a commercial software program, containing a proprietary source code that utilizes complex notions of high-level statistics such as a Markov Chain Monte Carlo engine. This process may generate fractions that contain astronomically high numbers presented in a bewildering format of conditional probability, all of which is based on underlying discretionary – and often unknowable – determinations.

The Government dismissively claims STRmix is "nothing new," but then lauds STRmix as a "significant development" that supposedly now can solve an extremely complex problem that was "not practically solvable" until recently.[34] Irrespective of these contradictory characterizations, the experts are still trying to grasp how to use this technology in situations involving complex mixtures. In fact, this very subject is the topic of intense debate even within

[32] Exhibit 12, William C. Thompson, "Are Juries Competent to Evaluate Statistical Evidence?" LAW & CONTEMPORARY PROBS, Vol 52: No 4 (1989), p 25 ("A juror who hears that the defendant and the perpetrator both have a blood type found in only 10 percent of the population, for example, may reason that there is only a 10 percent chance that the defendant would happen to have this blood type if he was innocent. The juror may then jump to the mistaken conclusion that there is therefore a 90 percent chance he is guilty.")

[33] Exhibit 13, Laurence H. Tribe, Trial by Mathematics: Precision and Ritual in the Legal Process, 84 HARV L REV 1329, 1334 (1971).

[34] ECF No. 52, p 7.

the Federal Government itself, as the position taken by the Government in its Response suggests.[35] In fact, the experts at the National Institute of Standards and Technology ("NIST") currently are engaged in a large-scale "scientific foundation review" of STRmix and other probabilistic genotyping systems to, in NIST's own words, "assess the reliability of forensic methods for analyzing DNA evidence that, if misapplied, could lead to innocent people being wrongly convicted."[36] The evidence at issue here is outside the boundaries of acceptable use at this time, because clear guideposts from the scientific community for using this technology reliably and effectively do not exist.

Finally, the problem of prejudice cannot be alleviated by vigorous cross-examination. As Government states in its Response, DNA is "a powerful tool" that "is ubiquitous to the point where it has infiltrated popular culture."[37] But as the PCAST Addendum states, "precisely because the conclusions are potentially so powerful and persuasive, the law requires scientific testimony be based on methods that are scientifically valid and reliable."[38] Validity and reliability of this method have not been demonstrated here. Furthermore, the impact of DNA evidence in a criminal trial is much greater when other non-DNA evidence in the case includes unreliable and inconsistent witness statements, which as the Government notes in its Response, is a circumstance present here.[39]

[35] ECF No. 52, p 18.

[36] Exhibit 11, "NIST To Assess the Reliability of Forensic Methods for Analyzing DNA Mixtures," Oct. 3, 2017, located at www.nist.gov/news-events/news/2017/10/nist-assess-reliability-forensic-methodsanalyzing-dna-mixtures.

[37] ECF No. 52, p 8.

[38] Exhibit 6, PCAST Report Addendum.

[39] ECF No. 52, pp 5-6.

WHEREFORE, the defendant, Daniel Gissantaner, respectfully requests that this Court grant his Motion to Exclude DNA Evidence.

Respectfully submitted,

SHARON A. TUREK
Federal Public Defender

Dated: February 23, 2018

/s/ Joanna C. Kloet
JOANNA C. KLOET
Assistant Federal Public Defender
50 Louis, NW, Suite 300
Grand Rapids, MI 49503
(616) 742-7420

Exhibit 1

PCAST Report
Excerpt

REPORT TO THE PRESIDENT
Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods

Executive Office of the President
President's Council of Advisors on
Science and Technology

September 2016

3. The Role of Scientific Validity in the Courts

The central focus of this report is the scientific validity of forensic-science evidence—more specifically, evidence from scientific methods for comparison of features (in, for example, DNA, latent fingerprints, bullet marks and other items). The reliability of methods for interpreting evidence is a fundamental consideration throughout science. Accordingly, every scientific field has a well-developed, domain-specific understanding of what scientific validity of methods entails.

The concept of scientific validity also plays an important role in the legal system. In particular, as noted in Chapter 1, the Federal Rules of Evidence require that expert testimony about forensic science must be the product of "reliable principles and methods" that have been "reliably applied . . . to the facts of the case."

This report explicates the scientific criteria for scientific validity in the case of forensic feature-comparison methods, for use both within the legal system and by those working to strengthen the scientific underpinnings of those disciplines. Before delving into that scientific explication, we provide in this chapter a very brief summary, aimed principally at scientists and lay readers, of the relevant legal background and terms, as well as the nature of this intersection between law and science.

3.1 Evolution of Admissibility Standards

Over the course of the 20th century, the legal system's approach for determining the admissibility of scientific evidence has evolved in response to advances in science. In 1923, in *Frye v. United States,*[81] the Court of Appeals for the District of Columbia considered the admissibility of testimony concerning results of a purported "lie detector," a systolic-blood- pressure deception test that was a precursor to the polygraph machine. After describing the device and its operation, the Court rejected the testimony, stating:

> *[W]hile courts will go a long way in admitting expert testimony deduced from a well-recognized scientific principle or discovery, the thing from which the deduction is made must be sufficiently established to have gained general acceptance in the particular field in which it belongs.*[82]

The court found that the systolic test had "not yet gained such standing and scientific recognition among physiological and psychological authorities," and was therefore inadmissible.

More than a half-century later, the Federal Rules of Evidence were enacted into law in 1975 to guide criminal and civil litigation in Federal courts. Rule 702, in its original form, stated that:

[81] *Frye v. United States*, 293 F. 1013 (D.C. Cir. 1923).
[82] Ibid., 1014.

> *If scientific, technical, or other specialized knowledge will assist the trier of fact to understand the evidence or to determine a fact in issue, a witness qualified as an expert by knowledge, skill, experience, training, or education, may testify thereto in the form of an opinion or otherwise.*[83]

There was considerable debate among litigants, judges, and legal scholars as to whether the rule embraced the *Frye* standard or established a new standard.[84] In 1993, the United States Supreme Court sought to resolve these questions in its landmark ruling in *Daubert v. Merrell Dow Pharmaceuticals*. In interpreting Rule 702, the *Daubert* Court held that the Federal Rules of Evidence superseded *Frye* as the standard for admissibility of expert evidence in Federal courts. The Court rejected "general acceptance" as the standard for admissibility and instead held that the admissibility of scientific expert testimony depended on its scientific reliability.

Where *Frye* told judges to defer to the judgment of the relevant expert community, *Daubert* assigned trial court judges the role of "gatekeepers" charged with ensuring that expert testimony "rests on reliable foundation."[85]

The Court stated that "the trial judge must determine . . . whether the reasoning or methodology underlying the testimony is scientifically valid."[86] It identified five factors that a judge should, among others, ordinarily consider in evaluating the validity of an underlying methodology. These factors are: (1) whether the theory or technique can be (and has been) tested; (2) whether the theory or technique has been subjected to peer review and publication; (3) the known or potential rate of error of a particular scientific technique; (4) the existence and maintenance of standards controlling the technique's operation; and (5) a scientific technique's degree of acceptance within a relevant scientific community.

The *Daubert* court also noted that judges evaluating proffers of expert scientific testimony should be mindful of other applicable rules, including:

- Rule 403, which permits the exclusion of relevant evidence "if its probative value is substantially outweighed by the danger of unfair prejudice, confusion of the issues, or misleading the jury…" (noting that expert evidence can be "both powerful and quite misleading because of the difficulty in evaluating it."); and
- Rule 706, which allows the court at its discretion to procure the assistance of an expert of its own choosing.[87]

[83] Act of January 2, 1975, Pub. Law No. 93-595, 88 Stat. 1926 (1975). See: federalevidence.com/pdf/FRE_Amendments/1975_Orig_Enact/1975-Pub.L._93-595_FRE.pdf.

[84] See: Giannelli, P.C. "The admissibility of novel scientific evidence: Frye v. United States, a half-century later." *Columbus Law Review*, Vol. 80, No. 6 (1980); McCabe, J. "DNA fingerprinting: The failings of Frye," *Norther Illinois University Law Review*, Vol. 16 (1996): 455-82; and Page, M., Taylor, J., and M. Blenkin. "Forensic identification science evidence since Daubert: Part II—judicial reasoning in decisions to exclude forensic identification evidence on grounds of reliability." *Journal of Forensic Sciences*, Vol. 56, No. 4 (2011): 913-7.

[85] *Daubert*, at 597.

[86] *Daubert*, at 580. See also, FN9 ("In a case involving scientific evidence, *evidentiary reliability* will be based on *scientific validity*." [emphasis in original]).

[87] *Daubert*, at 595, citing Weinstein, 138 F.R.D., at 632.

Congress amended Rule 702 in 2000 to make it more precise, and made further stylistic changes in 2011. In its current form, Rule 702 imposes four requirements:

> *A witness who is qualified as an expert by knowledge, skill, experience, training, or education may testify in the form of an opinion or otherwise if:*
> *(a) the expert's scientific, technical, or other specialized knowledge will help the trier of fact to understand the evidence or to determine a fact in issue;*
> *(b) the testimony is based on sufficient facts or data;*
> *(c) the testimony is the product of reliable principles and methods; and*
> *(d) the expert has reliably applied the principles and methods to the facts of the case.*

An Advisory Committee's Note to Rule 702 also specified a number of reliability factors that supplement the five factors enumerated in Daubert. Among those factors is "whether the field of expertise claimed by the expert is known to reach reliable results."[88,89]

Many states have adopted rules of evidence that track key aspects of these federal rules. Such rules are now the law in over half of the states, while other states continue to follow the Frye standard or variations of it.[90]

3.2 Foundational Validity and Validity as Applied

As described in *Daubert*, the legal system envisions an important conversation between law and science:

> *"The [judge's] inquiry envisioned by Rule 702 is, we emphasize, a flexible one. Its overarching subject is the scientific validity—and thus the evidentiary relevance and reliability—of the principles that underlie a proposed submission."*[91]

[88] See: Fed. R. Evid. 702 Advisory Committee note (2000). The following factors may be relevant under Rule 702: whether the underlying research was conducted independently of litigation; whether the expert unjustifiably extrapolated from an accepted premise to an unfounded conclusion; whether the expert has adequately accounted for obvious alternative explanations; whether the expert was as careful as she would be in her professional work outside of paid litigation; and *whether the field of expertise claimed by the expert is known to reach reliable results* [emphasis added].

[89] This note has been pointed to as support for efforts to challenge entire fields of forensic science, including fingerprints and hair comparisons. See: Giannelli, P.C. "The Supreme Court's 'Criminal' *Daubert* Cases." *Seton Hall Law Review,* Vol. 33 (2003): 1096.

[90] Even under the *Frye* formulation, the views of scientists about the meaning of reliability are relevant. *Frye* requires that a scientific technique or method must "have general acceptance" in the relevant scientific community to be admissible. As a scientific matter, the relevant scientific community for assessing the reliability of feature-comparison sciences includes metrologists (including statisticians) as well as other physical and life scientists from disciplines on which the specific methods are based. Importantly, the community is not limited to forensic scientists who practice the specific method. For example, the *Frye* court evaluated whether the proffered lie detector had gained "standing and scientific recognition among physiological and psychological authorities," rather than among lie detector experts. *Frye v. United States*, 293 F. 1013 (D.C. Cir. 1923).

[91] *Daubert*, at 594

Legal and scientific considerations thus both play important roles.

(1) The admissibility of expert testimony depends on a threshold test of, among other things, whether it meets certain *legal* standards embodied in Rule 702. These decisions about admissibility are exclusively the province of the courts.

(2) Yet, as noted above, the overarching subject of the judge's inquiry under Rule 702 is "scientific validity." It is the proper province of the scientific community to provide guidance concerning *scientific* standards for scientific validity.

PCAST does not opine here on the legal standards, but seeks only to clarify the scientific standards that underlie them. For complete clarity about our intent, we have adopted specific terms to refer to the *scientific* standards for two key types of scientific validity, which we mean to correspond, as scientific standards, to the legal standards in Rule 702 (c,d)):

(1) by "foundational validity," we mean the *scientific* standard corresponding to the legal standard of evidence being based on "reliable principles and methods," and

(2) by "validity as applied," we mean the *scientific* standard corresponding to the legal standard of an expert having "reliably applied the principles and methods."

In the next chapter, we turn to discussing the scientific standards for these concepts. We close this chapter by noting that answering the question of scientific validity in the forensic disciplines is important not just for the courts but also because it sets quality standards that ripple out throughout these disciplines—affecting practice and defining necessary research.

4. Scientific Criteria for Validity and Reliability of Forensic Feature-Comparison Methods

In this report, PCAST has chosen to focus on defining the validity and reliability of one specific area within forensic science: forensic feature-comparison methods. We have done so because it is both possible and important to do so for this particular class of methods.

- It is *possible* because feature comparison is a common scientific activity, and science has clear standards for determining whether such methods are reliable. In particular, feature-comparison methods belong squarely to the discipline of metrology—the science of measurement and its application.[92,93]

- It is *important* because it has become apparent, over the past decade, that faulty forensic feature comparison has led to numerous miscarriages of justice.[94] It has also been revealed that the problems

[92] International Vocabulary of Metrology – Basic and General Concepts and Associated Terms (VIM 3rd edition) JCGM 200 (2012).

[93] That forensic feature-comparison methods belong to the field of metrology is clear from the fact that NIST—whose mission is to assist the Nation by "advancing measurement science, standards and technology," and which is the world's leading metrological laboratory—is the home within the Federal government for research efforts on forensic science. NIST's programs include internal research, extramural research funding, conferences, and preparation of reference materials and standards. See: www.nist.gov/public_affairs/mission.cfm and www.nist.gov/forensics/index.cfm. Forensic feature-comparison methods involve determining whether two sets of features agree within a given measurement tolerance.

[94] DNA-based re-examination of past cases has led so far to the exonerations of 342 defendants, including 20 who had been sentenced to death, and to the identification of 147 real perpetrators. See: Innocence Project, "DNA Exonerations in the United States." www.innocenceproject.org/dna-exonerations-in-the-united-states. Reviews of these cases have revealed that roughly half relied in part on expert testimony that was based on methods that had not been subjected to meaningful scientific scrutiny or that included scientifically invalid claims of accuracy. See: Gross, S.R., and M. Shaffer. "Exonerations in the United States, 1989-2012." National Registry of Exonerations, (2012) available at: www.law.umich.edu/special/exoneration/Documents/exonerations_us_1989_2012_full_report.pdf; Garrett, B.L., and P.J. Neufeld. "Invalid forensic science testimony and wrongful convictions." *Virginia Law Review*, Vol. 91, No. 1 (2009): 1-97; National Research Council. *Strengthening Forensic Science in the United States: A Path Forward.* The National Academies Press. Washington DC. (2009): 42-3. The nature of the issues is illustrated by specific examples described in the materials cited: Levon Brooks and Kennedy Brewer, each convicted of separate child murders in the 1990s almost entirely on the basis of bitemark analysis testimony, spent more than 13 years in prison before DNA testing identified the actual perpetrator, who confessed to both crimes; Santae Tribble, convicted of murder after an FBI analyst testified that hair from a stocking mask linked Tribble to the crime and "matched in all microscopic characteristics," spent more than 20 years in prison before DNA testing revealed that none of the 13 hairs belonged to Tribble and that one came from a dog; Jimmy Ray Bromgard of Montana served 15 years in prison for rape before DNA testing showed that hairs collected from the victim's bed and reported as a match to Bromgard's could not have come from him; Stephan Cowans, convicted of shooting a Boston police officer after two fingerprint experts testified that a thumbprint left by the perpetrator was "unique and

> are not due simply to poor performance by a few practitioners, but rather to the fact that the reliability of many forensic feature-comparison methods has never been meaningfully evaluated.[95]

Compared to many types of expert testimony, testimony based on forensic feature-comparison methods poses unique dangers of misleading jurors for two reasons:

- The vast majority of jurors have no independent ability to interpret the probative value of results based on the detection, comparison, and frequency of scientific evidence. If matching halves of a ransom note were found at a crime scene and at a defendant's home, jurors could rely on their own experiences to assess how unlikely it is that two torn scraps would match if they were not in fact from a single original note. If a witness were to describe a perpetrator as "tall and bushy haired," jurors could make a reasonable judgment of how many people might match the description. But, if an expert witness were to say that, in two DNA samples, the third exon of the *DYNC1H1* gene is precisely 174 nucleotides in length, most jurors would have no way to know if they should be impressed by the coincidence; they would be completely dependent on expert statements garbed in the mantle of science. (As it happens, they should not be impressed by the preceding statement: At the DNA locus cited, more than 99.9 percent of people have a fragment of the indicated size.[96])

- The potential prejudicial impact is unusually high, because jurors are likely to overestimate the probative value of a "match" between samples. Indeed, the DOJ itself historically overestimated the probative value of matches in its longstanding contention, now acknowledged to be inappropriate, that latent fingerprint analysis was "infallible."[97] Similarly, a former head of the FBI's fingerprint unit testified that the FBI had "an error rate of one per every 11 million cases."[98] In an online experiment, researchers asked mock jurors to estimate the frequency that a qualified, experienced forensic scientist would mistakenly conclude that two samples of specified types came from the same person when they actually came from two different people. The mock jurors believed such errors are likely to occur about 1 in 5.5 million for fingerprint analysis comparison; 1 in 1 million for bitemark comparison; 1 in 1 million for hair comparison; and 1 in 100 thousand for handwriting comparison.[99] While precise error rates are not known for most of these techniques, all indications point to the actual error rates being orders of magnitude higher. For example, the FBI's own studies of latent fingerprint analysis point to error rates in the range of one in several hundred.[100] (Because the term "match" is likely to imply an

identical," spent more than 5 years in prison before DNA testing on multiple items of evidence excluded him as the perpetrator; and Steven Barnes of upstate New York served 20 years in prison for a rape and murder he did not commit after a criminalist testified that a photographic overlay of fabric from the victim's jeans and an imprint on Barnes' truck showed patterns that were "similar" and hairs collected from the truck were similar to the victim's hairs.

[95] See: Chapter 5.

[96] See: ExAC database: exac.broadinstitute.org/gene/ENSG00000197102.

[97] See: www.justice.gov/olp/file/861906/download.

[98] *U.S. v. Baines* 573 F.3d 979 (2009) at 984.

[99] Koehler, J.J. "Intuitive error rate estimates for the forensic sciences." (August 2, 2016). Available at papers.ssrn.com/sol3/papers.cfm?abstract_id=2817443 .

[100] See: Section 5.4.

inappropriately high probative value, a more neutral term should be used for an examiner's belief that two samples come from the same source. We suggest the term "*proposed* identification" to appropriately convey the examiner's conclusion, along with the possibility that it might be wrong. We will use this term throughout this report.)

This chapter lays out PCAST's conclusions concerning the scientific criteria for scientific validity. The conclusions are based on the fundamental principles of the "scientific method"—applicable throughout science—that valid scientific knowledge can *only* be gained through *empirical* testing of specific propositions.[101] PCAST's conclusions in the chapter might be briefly summarized as follows:

> *Scientific validity and reliability require that a method has been subjected to empirical testing, under conditions appropriate to its intended use, that provides valid estimates of how often the method reaches an incorrect conclusion. For subjective feature-comparison methods, appropriately designed black-box studies are required, in which many examiners render decisions about many independent tests (typically, involving "questioned" samples and one or more "known" samples) and the error rates are determined. Without appropriate estimates of accuracy, an examiner's statement that two samples are similar—or even indistinguishable—is scientifically meaningless: it has no probative value, and considerable potential for prejudicial impact. Nothing—not training, personal experience nor professional practices—can substitute for adequate empirical demonstration of accuracy.*

The chapter is organized as follows:

- The first section describes the distinction between two fundamentally different types of feature-comparison methods: objective methods and subjective methods.

- The next five sections discuss the scientific criteria for the two types of scientific validity: foundational validity and validity as applied.

- The final two sections discuss views held in the forensic community.

4.1 Feature-Comparison Methods: Objective and Subjective Methods

A forensic feature-comparison method is a procedure by which an examiner seeks to determine whether an evidentiary sample (e.g., from a crime scene) is or is not associated with a source sample (e.g., from a suspect)[102] based on similar features. The evidentiary sample might be DNA, hair, fingerprints, bitemarks, toolmarks, bullets, tire tracks, voiceprints, visual images, and so on. The source sample would be biological material or an item (tool, gun, shoe, or tire) associated with the suspect.

[101] For example, the Oxford Online Dictionary defines the scientific method as "a method or procedure that has characterized the natural sciences since the 17th century, consisting in systematic observation, measurement, and experimentation, and the formulation, testing, and modification of hypotheses." "Scientific method" *Oxford Dictionaries Online*. Oxford University Press (accessed on August 19, 2016).

[102] A "source sample" refers to a specific individual or object (e.g., a tire or gun).

Feature-comparison methods may be classified as either objective or subjective. By objective feature-comparison methods, we mean methods consisting of procedures that are each defined with enough standardized and quantifiable detail that they can be performed by either an automated system or human examiners exercising little or no judgment. By subjective methods, we mean methods including key procedures that involve significant human judgment—for example, about which features to select or how to determine whether the features are sufficiently similar to be called a proposed identification.

Objective methods are, in general, preferable to subjective methods. Analyses that depend on human judgment (rather than a quantitative measure of similarity) are obviously more susceptible to human error, bias, and performance variability across examiners.[103] In contrast, objective, quantified methods tend to yield greater accuracy, repeatability and reliability, including reducing variation in results among examiners. Subjective methods can evolve into or be replaced by objective methods.[104]

4.2 Foundational Validity: Requirement for Empirical Studies

For a metrological method to be scientifically valid and reliable, the procedures that comprise it must be shown, based on empirical studies, to be *repeatable*, *reproducible*, and *accurate*, at levels that have been measured and are appropriate to the intended application.[105,106]

BOX 2. Definition of key terms

By "repeatable," we mean that, with known probability, an examiner obtains the same result, when analyzing samples from the same sources.

By "reproducible," we mean that, with known probability, different examiners obtain the same result, when analyzing the same samples.

By "accurate," we mean that, with known probabilities, an examiner obtains correct results both (1) for samples from the same source (true positives) and (2) for samples from different sources (true negatives).

By "reliability," we mean repeatability, reproducibility, and accuracy.[107]

[103] Dror, I.E. "A hierarchy of expert performance." *Journal of Applied Research in Memory and Cognition*, Vol. 5 (2016): 121-127.

[104] For example, before the development of objective tests for intoxication, courts had to rely exclusively on the testimony of police officers and others who in turn relied on behavioral indications of drunkenness and the presence of alcohol on the breath. The development of objective chemical tests drove a change from subjective to objective standards.

[105] National Physical Laboratory. "A Beginner's Guide to Measurement." (2010) available at: www.npl.co.uk/upload/pdf/NPL-Beginners-Guide-to-Measurement.pdf; Pavese, F. "An Introduction to Data Modelling Principles in Metrology and Testing." in *Data Modeling for Metrology and Testing in Measurement Science*, Pavese, F. and A.B. Forbes (Eds.) Birkhäuser (2009).

[106] Feature-comparison methods that get the wrong answer too often have, by definition, low probative value. As discussed above, the prejudicial impact will thus likely to outweigh the probative value.

[107] We note that "reliability" also has a narrow meaning within the field of statistics referring to "consistency"—that is, the extent to which a method produces the same result, regardless of whether the result is accurate. This is not the sense in which "reliability" is used in this report, or in the law.

By "scientific validity," we mean that a method has shown, based on empirical studies, to be reliable with levels of repeatability, reproducibility, and accuracy that are appropriate to the intended application.

By an "empirical study," we mean test in which a method has been used to analyze a large number of independent sets of samples, similar in relevant aspects to those encountered in casework, in order to estimate the method's repeatability, reproducibility, and accuracy.

By a "black-box study," we mean an empirical study that assesses a subjective method by having examiners analyze samples and render opinions about the origin or similarity of samples.

The method need not be perfect, but it is clearly *essential* that its accuracy has been measured based on appropriate empirical testing and is high enough to be appropriate to the application. Without an appropriate estimate of its accuracy, a metrological method is useless—because one has no idea how to interpret its results. The importance of knowing a method's accuracy was emphasized by the 2009 NRC report on forensic science and by a 2010 NRC report on biometric technologies.[108]

To meet the scientific criteria of foundational validity, two key elements are required:

(1) a reproducible and consistent procedure for (a) identifying features within evidence samples; (b) comparing the features in two samples; and (c) determining, based on the similarity between the features in two samples, whether the samples should be declared to be a proposed identification ("matching rule").

(2) empirical measurements, from multiple independent studies, of (a) the method's false positive rate—that is, the probability it declares a proposed identification between samples that actually come from *different* sources and (b) the method's sensitivity—that is, probability that it declares a proposed identification between samples that actually come from the *same* source.

We discuss these elements in turn.

Reproducible and Consistent Procedures

For a method to be objective, *each* of the three steps (feature identification, feature comparison, and matching rule) should be precisely defined, reproducible and consistent. Forensic examiners should identify relevant features in the same way and obtain the same result. They should compare features in the same quantitative manner. To declare a proposed identification, they should calculate whether the features in an evidentiary sample and the features in a sample from a suspected source lie within a pre-specified measurement tolerance

[108] "Biometric recognition is an inherently probabilistic endeavor…Consequently, even when the technology and the system it is embedded in are behaving as designed, there is inevitable uncertainty and risk of error." National Research Council, *"Biometric Recognition: Challenges and Opportunities."* The National Academies Press. Washington DC. (2010): viii-ix.

(matching rule).[109] For an objective method, one can establish the foundational validity of each of the individual steps by measuring its accuracy, reproducibility, and consistency.

For subjective methods, procedures must still be carefully defined—but they involve substantial human judgment. For example, different examiners may recognize or focus on different features, may attach different importance to the same features, and may have different criteria for declaring proposed identifications. Because the procedures for feature identification, the matching rule, and frequency determinations about features are not objectively specified, the overall procedure must be treated as a kind of "black box" inside the examiner's head.

Subjective methods require careful scrutiny, more generally, their heavy reliance on human judgment means that they are especially vulnerable to human error, inconsistency across examiners, and cognitive bias. In the forensic feature-comparison disciplines, cognitive bias includes the phenomena that, in certain settings, humans (1) may tend naturally to focus on similarities between samples and discount differences and (2) may also be influenced by extraneous information and external pressures about a case.[110] (The latter issues are illustrated by the FBI's misidentification of a latent fingerprint in the Madrid training bombing, discussed on p.9.)

Since the black box in the examiner's head cannot be examined directly for its foundational basis in science, the foundational validity of subjective methods can be established *only* through empirical studies of examiner's performance to determine whether they can provide accurate answers; such studies are referred to as "black-box" studies (Box 2). In black-box studies, many examiners are presented with many independent comparison problems—typically, involving "questioned" samples and one or more "known" samples—and asked to declare whether the questioned samples came from the same source as one of the known samples.[111] The researchers then determine how often examiners reach erroneous conclusions.

[109] If a source is declared *not* to share the same features, it is "excluded" by the test. The matching rule should be chosen carefully. If the "matching rule" is chosen to be too strict, samples that actually come from the same source will be declared a non-match (false negative). If it is too lax, then the method will not have much discriminatory power because the random match probability will be too high (false positive).

[110] See, for example: Boroditsky, L. "Comparison and the development of knowledge." *Cognition*, Vol. 102 (2007): 118-128; Hassin, R. "Making features similar: comparison processes affect perception." *Psychonomic Bulletin & Review*, Vol. 8 (2001): 728–31; Medin, D.L., Goldstone, R.L., and D. Gentner. "Respects for similarity." *Psychological Review*, Vol. 100 (1993): 254–78; Tversky, A. "Features of similarity." *Psychological Review*, Vol. 84 (1977): 327–52; Kim, J., Novemsky, N., and R. Dhar. "Adding small differences can increase similarity and choice." *Psychological Science*, Vol. 24 (2012): 225–9; Larkey, L.B., and A.B. Markman. "Processes of similarity judgment." *Cognitive Science*, Vol. 29 (2005): 1061–76; Medin, D.L., Goldstone, R.L., and A.B. Markman. "Comparison and choice: Relations between similarity processes and decision processes." *Psychonomic Bulletin and Review*, Vol. 2 (1995): 1–19; Goldstone, R. L. "The role of similarity in categorization: Providing a groundwork." *Cognition*, Vol. 52 (1994): 125–57; Nosofsky, R. M. "Attention, similarity, and the identification-categorization relation." *Journal of Experimental Psychology, General*, Vol. 115 (1986): 39–57.

[111] Answers may be expressed in such terms as "match/no match/inconclusive" or "identification/exclusion/inconclusive."

As an excellent example, the FBI recently conducted a black-box study of latent fingerprint analysis, involving 169 examiners and 744 fingerprint pairs, and published the results of the study in a leading scientific journal.[112]

(Some forensic scientists have cautioned that too much attention to the subjective aspects of forensic methods—such as studies of cognitive bias and black-box studies—might distract from the goal of improving knowledge about the objective features of the forensic evidence and developing truly objective methods.[113] Others have noted that this is not currently a problem, because current efforts and funding to address the challenges associated with subjective forensic methods are very limited.[114])

Empirical Measurements of Accuracy

It is necessary to have appropriate empirical measurements of a method's false positive rate and the method's sensitivity. As explained in Appendix A, it is necessary to know these two measures to assess the probative value of a method.

The false positive rate is the probability that the method declares a proposed identification between samples that actually come from *different* sources. For example, a false positive rate of 5 percent means that two samples from *different* sources will (due to limitations of the method) be incorrectly declared to come from the same source 5 percent of the time. (The quantity equal to one minus the false positive rate—95 percent, in the example—is referred to as the specificity.)

The method's sensitivity is the probability that the method declares a proposed identification between samples that actually come from the *same* source. For example, a sensitivity of 90 percent means two samples from the same source will be declared to come from the same source 90 percent of the time, and declared to come from different sources 10 percent of the time. (The latter quantity is referred to as the false negative rate.)

The false positive rate is especially important because false positive results can lead directly to wrongful convictions.[115] In some circumstances, it may be possible to estimate a false positive rate related to specific features of the evidence in the case. (For example, the random match probability calculated in DNA analysis depends in part on the specific genotype seen in an evidentiary sample. The false positive rate for latent fingerprint analysis may depend on the quality of the latent print.) For other feature-comparison methods, it may be only possible to make an overall estimate of the average false positive rate across samples.

For objective methods, the false positive rate is composed of two distinguishable sources—coincidental matches (where samples from different sources nonetheless have *features* that fall within the tolerance of the objective matching rule) and human/technical failures (where samples have features that fall outside the matching rule, but where a proposed identification was nonetheless declared due to a human or technical failure). For

[112] Ulery, B.T., Hicklin, R.A., Buscaglia, J., and M.A. Roberts. "Accuracy and reliability of forensic latent fingerprint decisions." *Proceedings of the National Academy of Sciences*, Vol. 108, No. 19 (2011): 7733-8.
[113] Champod, C. "Research focused mainly on bias will paralyse forensic science." *Science & Justice*, Vol. 54 (2014): 107–9.
[114] Risinger, D.M., Thompson, W.C., Jamieson, A., Koppl, R., Kornfield, I., Krane, D., Mnookin, J.L., Rosenthal, R., Saks, M.J., and S.L. Zabell. "Regarding Champod, editorial: "Research focused mainly on bias will paralyse forensic science." *Science and Justice*, Vol. 54 (2014):508-9.
[115] See footnote 94, p. 44. Under some circumstances, false-negative results can contribute to wrongful convictions as well.

objective methods where the probability of coincidental match is very low (such as DNA analysis), the false positive rate in application in a given case will be dominated by the rate of human/technical failures—which may well be hundreds of times larger.

For subjective methods, both types of error—coincidental matches and human/technical failures—occur as well, but, without an objective "matching rule," the two sources cannot be distinguished. In establishing foundational validity, it is thus essential to perform black-box studies that empirically measure the overall error rate across many examiners. (See Box 3 concerning the word "error.")

BOX 3. The meanings of "error"

The term "error" has differing meanings in science and law, which can lead to confusion. In legal settings, the term "error" often implies fault—e.g., that a person has made a mistake that could have been avoided if he or she had properly followed correct procedures or a machine has given an erroneous result that could have been avoided it if had been properly calibrated. In science, the term "error" also includes the situation in which the procedure itself, when properly applied, does not yield the correct answer owing to chance occurrence.

When one applies a forensic feature-comparison method with the goal of assessing whether two samples did or did not come from the same source, coincidental matches and human/technical failures are both regarded, from a statistical point of view, as "errors" because both can lead to incorrect conclusions.

Studies designed to estimate a method's false positive rate and sensitivity are necessarily conducted using only a finite number of samples. As a consequence, they cannot provide "exact" values for these quantities (and should not claim to do so), but only "confidence intervals," whose bounds reflect, respectively, the range of values that are reasonably compatible with the results. When reporting a false positive rate to a jury, it is scientifically important to state the "upper 95 percent one-sided confidence bound" to reflect the fact that the actual false positive rate could reasonably be as high as this value.[116] (For more information, see Appendix A.)

Studies often categorize their results as being conclusive (e.g., identification or exclusion) or inconclusive (no determination made).[117] When reporting a false positive rate to a jury, it is scientifically important to calculate the rate based on the proportion of *conclusive* examinations, rather than just the proportion of all examinations. This is appropriate because evidence used against a defendant will typically be based on *conclusive*, rather than inconclusive, examinations. To illustrate the point, consider an extreme case in which a method had been

[116] The upper confidence bound properly incorporates the precision of the estimate based on the sample size. For example, if a study found no errors in 100 tests, it would be misleading to tell a jury that the error rate was 0 percent. In fact, if the tests are independent, the upper 95 percent confidence bound for the true error rate is 3.0 percent. Accordingly a jury should be told that the error rate could be as high as 3.0 percent (that is, 1 in 33). The true error rate could be higher, but with rather small probability (less than 5 percent). If the study were much smaller, the upper 95 percent confidence limit would be higher. For a study that found no errors in 10 tests, the upper 95 percent confidence bound is 26 percent—that is, the actual false positive rate could be roughly 1 in 4 (see Appendix A).
[117] See: Chapter 5.

tested 1000 times and found to yield 990 inconclusive results, 10 false positives, and no correct results. It would be misleading to report that the false positive rate was 1 percent (10/1000 examinations). Rather, one should report that 100 percent of the conclusive results were false positives (10/10 examinations).

Whereas exploratory scientific studies may take many forms, scientific *validation* studies—intended to assess the validity and reliability of a metrological method for a particular forensic feature-comparison application—must satisfy a number of criteria, which are described in Box 4.

BOX 4. Key criteria for validation studies to establish foundational validity

Scientific validation studies—intended to assess the validity and reliability of a metrological method for a particular forensic feature-comparison application—must satisfy a number of criteria.

(1) The studies must involve a sufficiently large number of examiners and must be based on sufficiently *large* collections of *known* and *representative* samples from *relevant* populations to reflect the range of features or combinations of features that will occur in the application. In particular, the sample collections should be:

(a) representative of the quality of evidentiary samples seen in real cases. (For example, if a method is to be used on distorted, partial, latent fingerprints, one must determine the *random match probability*—that is, the probability that the match occurred by chance—for distorted, partial, latent fingerprints; the random match probability for full scanned fingerprints, or even very high quality latent prints would not be relevant.)

(b) chosen from populations relevant to real cases. For example, for features in biological samples, the false positive rate should be determined for the overall US population and for major ethnic groups, as is done with DNA analysis.

(c) large enough to provide appropriate estimates of the error rates.

(2) The empirical studies should be conducted so that neither the examiner nor those with whom the examiner interacts have any information about the correct answer.

(3) The study design and analysis framework should be specified in advance. In validation studies, it is inappropriate to modify the protocol afterwards based on the results.[118]

[118] The analogous situation in medicine is a clinical trial to test the safety and efficacy of a drug for a particular application. In the design of clinical trials, FDA requires that criteria for analysis must be pre-specified and notes that *post hoc* changes to the analysis compromise the validity of the study. See: FDA Guidance: "Adaptive Designs for Medical Device Clinical Studies" (2016) Available at: www.fda.gov/downloads/medicaldevices/deviceregulationandguidance/guidancedocuments/ucm446729.pdf; Alosh, M., Fritsch, K., Huque, M., Mahjoob, K., Pennello, G., Rothmann, M., Russek-Cohen, E., Smith, F., Wilson, S., and L. Yue. "Statistical considerations on subgroup analysis in clinical trials." *Statistics in Biopharmaceutical Research*, Vol. 7 (2015): 286-303; FDA Guidance: "Design Considerations for Pivotal Clinical Investigations for Medical Devices" (2013) (available at:

(4) The empirical studies should be conducted or overseen by individuals or organizations that have no stake in the outcome of the studies.[119]

(5) Data, software and results from validation studies should be available to allow other scientists to review the conclusions.

(6) To ensure that conclusions are reproducible and robust, there should be multiple studies by separate groups reaching similar conclusions.

An empirical measurement of error rates is not simply a desirable feature; it is *essential* for determining whether a method is foundationally valid. In science, a testing procedure—such as testing whether a person is pregnant or whether water is contaminated—is not considered valid until its reliability has been *empirically* measured. For example, we need to know how often the pregnancy test declares a pregnancy when there is none, and *vice versa*. The same scientific principles apply no less to forensic tests, which may contribute to a defendant losing his life or liberty.

Importantly, error rates cannot be inferred from casework, but rather must be determined based on samples where the correct answer is known. For example, the former head of the FBI's fingerprint unit testified that the FBI had "an error rate of one per every 11 million cases" based on the fact that the agency was known to have made only one mistake over the past 11 years, during which time it had made 11 million identifications.[120] The fallacy is obvious: the expert simply *assumed without evidence* that every error in casework had come to light.

Why is it essential to know a method's false positive rate and sensitivity? Because without appropriate empirical measurement of a method's accuracy, the fact that two samples in a particular case show similar features has *no probative value*—and, as noted above, it may have considerable prejudicial impact because juries will likely incorrectly attach meaning to the observation.[121]

www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm373750.htm); FDA Guidance for Industry: E9 Statistical Principles for Clinical Trials (September 1998) (available at: www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm073137.pdf); Pocock, S.J. Clinical trials: a practical approach. Wiley, Chichester (1983).

[119] In the setting of clinical trials, the sponsor of the trial (a pharmaceutical, device or biotech company or, in some cases, an academic institutions) funds and initiates the study, but the trial is conducted by individuals who are independent of the sponsor (often, academic physicians), in order to ensure the reliability of the data generated by the study and minimize the potential for bias. See, for example, 21 C.F.R. § 312.3 and 21 C.F.R. § 54.4(a).

[120] *U.S. v. Baines* 573 F.3d 979 (2009) at 984.

[121] Under Fed. R. Evid., Rule 403, evidence should be excluded "if its probative value is substantially outweighed by the danger of unfair prejudice."

The absolute need, from a scientific perspective, for empirical data is elegantly expressed in an analogy by U.S. District Judge John Potter in his opinion in *U.S. v. Yee (1991),* an early case on the use of DNA analysis:

> *Without the probability assessment, the jury does not know what to make of the fact that the patterns match: the jury does not know whether the patterns are as common as pictures with two eyes, or as unique as the Mona Lisa.*[122,123]

4.3 Foundational Validity: Requirement for Scientifically Valid Testimony

It should be obvious—but it bears emphasizing—that once a method has been established as foundationally valid based on appropriate empirical studies, claims about the method's accuracy and the probative value of proposed identifications, in order to be valid, must be based on such empirical studies. *Statements claiming or implying greater certainty than demonstrated by empirical evidence are scientifically invalid*. Forensic examiners should therefore report findings of a proposed identification with clarity and restraint, explaining in each case that the fact that two samples satisfy a method's criteria for a proposed match does not necessarily imply that the samples come from a common source. If the false positive rate of a method has been found to be 1 in 50, experts should not imply that the method is able to produce results at a higher accuracy.

Troublingly, expert witnesses sometimes go beyond the empirical evidence about the frequency of features—even to the extent of claiming or implying that a sample came from a specific source with near-certainty or even absolute certainty, despite having no scientific basis for such opinions.[124] From the standpoint of scientific validity, experts should never be permitted to state or imply in court that they can draw conclusions with certainty or near-certainty (such as "zero," "vanishingly small," "essentially zero," "negligible," "minimal," or "microscopic" error rates; "100 percent certainty" or "to a reasonable degree of scientific certainty;" or identification "to the exclusion of all other sources."[125]

The scientific inappropriateness of such testimony is aptly captured by an analogy by District of Columbia Court of Appeals Judge Catharine Easterly in her concurring opinion in *Williams v. United States*, a case in which an examiner testified that markings on certain bullets were unique to a gun recovered from a defendant's apartment:

[122] *U.S. v. Yee,* 134 F.R.D. 161 (N.D. Ohio 1991).

[123] Some courts have ruled that there is no harm in admitting feature-comparison evidence on the grounds that jurors can see the features with their own eyes and decide for themselves about whether features are shared. *U.S. v. Yee* shows why this reasoning is fallacious: jurors have no way to know how often two different samples would share features, and to what level of specificity.

[124] As noted above, the long history of exaggerated claims for the accuracy of forensic methods includes the DOJ's own prior statement that latent fingerprint analysis was "infallible," which the DOJ has judged to have been inappropriate. www.justice.gov/olp/file/861906/download.

[125] Cole, S.A. "Grandfathering evidence: Fingerprint admissibility rulings from Jennings to Llera Plaza and back again." *41 American Criminal Law Review, 1189* (2004). See also: National Research Council. *Strengthening Forensic Science in the United States: A Path Forward.* The National Academies Press. Washington DC. (NRC Report, 2009): 87, 104, and 143.

> *As matters currently stand, a certainty statement regarding toolmark pattern matching has the same probative value as the vision of a psychic: it reflects nothing more than the individual's foundationless faith in what he believes to be true. This is not evidence on which we can in good conscience rely, particularly in criminal cases, where we demand proof—real proof—beyond a reasonable doubt, precisely because the stakes are so high.*[126]

In science, assertions that a metrological method is more accurate than has been empirically demonstrated are rightly regarded as mere speculation, not valid conclusions that merit credence.

4.4 Neither Experience nor Professional Practices Can Substitute for Foundational Validity

In some settings, an expert may be scientifically capable of rendering judgments based primarily on his or her "experience" and "judgment." Based on experience, a surgeon might be scientifically qualified to offer a judgment about whether another doctor acted appropriately in the operating theater or a psychiatrist might be scientifically qualified to offer a judgment about whether a defendant is mentally competent to assist in his or her defense.

By contrast, "experience" or "judgment" cannot be used to establish the scientific validity and reliability of a metrological method, such as a forensic feature-comparison method. The frequency with which a particular pattern or set of features will be observed in different samples, which is an essential element in drawing conclusions, is not a matter of "judgment." It is an empirical matter for which only empirical evidence is relevant. Moreover, a forensic examiner's "experience" from extensive casework is not informative—because the "right answers" are not typically known in casework and thus examiners cannot accurately know how often they erroneously declare matches and cannot readily hone their accuracy by learning from their mistakes in the course of casework.

Importantly, good professional practices—such as the existence of professional societies, certification programs, accreditation programs, peer-reviewed articles, standardized protocols, proficiency testing, and codes of ethics—cannot substitute for actual evidence of scientific validity and reliability.[127]

Similarly, an expert's expression of *confidence* based on personal professional experience or expressions of *consensus* among practitioners about the accuracy of their field is no substitute for error rates estimated from relevant studies. For a method to be *reliable*, empirical evidence of validity, as described above, is required.

Finally, the points above underscore that scientific validity of a method must be assessed within the framework of the broader scientific field of which it is a part (e.g., measurement science in the case of feature-comparison methods). The fact that bitemark examiners defend the validity of bitemark examination means little.

[126] *Williams v. United States,* DC Court of Appeals, decided January 21, 2016, (Easterly, concurring).

[127] For example, both scientific and pseudoscientific disciplines employ such practices.

4.5 Validity as Applied: Key Elements

Foundational validity means that a method can, *in principle,* be reliable. Validity as applied means that the method has been reliably applied *in practice.* It is the *scientific* concept we mean to correspond to the legal requirement, in Rule 702(d), that an expert "has reliably applied the principles and methods to the facts of the case."

From a scientific standpoint, certain criteria are essential to establish that a forensic practitioner has reliably applied a method to the facts of a case. These elements are described in Box 5.

BOX 5. Key criteria for validity as applied

(1) The forensic examiner must have been shown to be *capable* of reliably applying the method and must *actually* have done so. Demonstrating that an examiner is *capable* of reliably applying the method is crucial—especially for subjective methods, in which human judgment plays a central role. From a scientific standpoint, the ability to apply a method reliably can be demonstrated only through empirical testing that measures how often the expert reaches the correct answer. (Proficiency testing is discussed more extensively on p. 57-59.) Determining whether an examiner has *actually* reliably applied the method requires that the procedures actually used in the case, the results obtained, and the laboratory notes be made available for scientific review by others.

(2) Assertions about the probability of the observed features occurring by chance must be scientifically valid.

(a) The forensic examiner should report the overall false positive rate and sensitivity for the method established in the studies of foundational validity and should demonstrate that the samples used in the foundational studies are relevant to the facts of the case.[128]

(b) Where applicable, the examiner should report the random match probability based on the specific features observed in the case.

(c) An expert should not make claims or implications that go beyond the empirical evidence and the applications of valid statistical principles to that evidence.

[128] For example, for DNA analysis, the frequency of genetic variants is known to vary among ethnic groups; it is thus important that the sample collection reflect relevant ethnic groups to the case at hand. For latent fingerprints, the risk of falsely declaring an identification may be higher when latent fingerprints are of lower quality; so, to be relevant, the sample collections used to estimate accuracy should be based on latent fingerprints comparable in quality and completeness to the case at hand.

4.6 Validity as Applied: Proficiency Testing

Even when a method is foundationally valid, there are many reasons why examiners may not always get the right result.[129] As discussed above, the *only* way to establish scientifically that an examiner is capable of applying a foundationally valid method is through appropriate empirical testing to measure how often the examiner gets the correct answer.

Such empirical testing is often referred to as "proficiency testing." We note that term "proficiency testing" is sometimes used to refer to many different other types of testing—such as (1) tests to determine whether a practitioner reliably follows the steps laid out in a protocol, without assessing the *accuracy* of their conclusions, and (2) practice exercises that help practitioners improve their skills by highlighting their errors, without accurately reflect the circumstances of actual casework.

In this report, we use the term proficiency testing to mean ongoing empirical tests to "evaluate the capability and performance of analysts."[130, 131, 132]

Proficiency testing should be performed under conditions that are representative of casework and on samples, for which the true answer is known, that are representative of the full range of sample types and quality likely to be encountered in casework in the intended application. (For example, the fact that an examiner passes a proficiency test involving DNA analysis of simple, single-source samples does not demonstrate that they are capable of DNA analysis of complex mixtures of the sort encountered in casework; see p. 76-81.)

To ensure integrity, proficiency testing should be overseen by a disinterested third party that has no institutional or financial incentive to skew performance. We note that testing services have stated that forensic community prefers that tests not be too challenging.[133]

[129] J.J. Koehler has enumerated a number of possible problems that could, in principle, occur: features may be mismeasured; samples may be interchanged, mislabeled, miscoded, altered, or contaminated; equipment may be miscalibrated; technical glitches and failures may occur without warning and without being noticed; and results may be misread, misinterpreted, misrecorded, mislabeled, mixed up, misplaced, or discarded. Koehler, J.J. "Forensics or fauxrensics? Ascertaining accuracy in the forensic sciences." papers.ssrn.com/sol3/papers.cfm?abstract_id=2773255 (accessed June 28, 2016).

[130] ASCLD/LAB Supplemental Requirements for Accreditation of Forensic Testing Laboratories. des.wa.gov/SiteCollectionDocuments/About/1063/RFP/Add7_Item4ASCLD.pdf.

[131] We note that proficiency testing is not intended to estimate the inherent error rates of a method; these rates should be assessed from foundational validity studies.

[132] Proficiency testing should also be distinguished from "competency testing," which is "the evaluation of a person's knowledge and ability prior to performing independent work in forensic casework." des.wa.gov/SiteCollectionDocuments/About/1063/RFP/Add7_Item4ASCLD.pdf.

[133] Christopher Czyryca, the president of Collaborative Testing Services, Inc., the leading proficiency testing firm in the U.S., has publicly stated that "Easy tests are favored by the community." August 2015 meeting of the National Commission on Forensic Science, a presentation at the Accreditation and Proficiency Testing Subcommittee. www.justice.gov/ncfs/file/761061/download.

As noted previously, false positive rates consist of both coincidental match rates and technical/human failure rates. For some technologies (such as DNA analysis), the latter may be hundreds of times higher than the former.

Proficiency testing is especially critical for subjective methods: because the procedure is not based solely on objective criteria but relies on human judgment, it is inherently vulnerable to error and inter-examiner variability. Each examiner should be tested, because empirical studies have noted considerable differences in accuracy across examiners.[134,135]

The test problems used in proficiency tests should be publicly released after the test is completed, to enable scientists to assess the appropriateness and adequacy of the test for their intended purpose.

Finally, proficiency testing should *ideally* be conducted in a 'test-blind' manner—that is, with samples inserted into the flow of casework such that examiners do not know that they are being tested. (For example, the Transportation Security Administration conducts blind tests by sending weapons and explosives inside luggage through screening checkpoints to see how often TSA screeners detect them.) It has been established in many fields (including latent fingerprint analysis) that, when individuals are aware that they are being tested, they perform differently than they do in the course of their daily work (referred to as the "Hawthorne Effect").[136,137]

While test-blind proficiency testing is ideal, there is disagreement in the forensic community about its feasibility in all settings. On the one hand, laboratories vary considerably as to the type of cases they receive, how evidence is managed and processed, and what information is provided to an analyst about the evidence or the case in question. Accordingly, blinded, inter-laboratory proficiency tests may be difficult to design and

[134] For example, a 2011 study on latent fingerprint decisions observed that examiners frequently differed on whether fingerprints were suitable for reaching a conclusion. Ulery, B.T., Hicklin, R.A., Buscaglia, J., and M.A. Roberts. "Accuracy and reliability of forensic latent fingerprint decisions." *Proceedings of the National Academy of Sciences*, Vol. 108, No. 19 (2011): 7733-8.

[135] It is not sufficient to point to proficiency testing on volunteers in a laboratory, because better performing examiners are more likely to participate. Koehler, J.J. "Forensics or fauxrensics? Ascertaining accuracy in the forensic sciences." papers.ssrn.com/sol3/papers.cfm?abstract_id=2773255 (accessed June 28, 2016).

[136] Concerning the Hawthorne effect, see, for example: Bracht, G.H., and G.V. Glass. "The external validity of experiments." *American Educational Research Journal,* Vol. 5, No. 4 (1968): 437-74; Weech, T.L. and H. Goldhor. "Obtrusive versus unobtrusive evaluation of reference service in five Illinois public libraries: A pilot study." *Library Quarterly: Information, Community, Policy*, Vol. 52, No. 4 (1982): 305-24; Bouchet, C., Guillemin, F., and S. Braincon. "Nonspecific effects in longitudinal studies: impact on quality of life measures." *Journal of Clinical Epidemiology,* Vol. 49, No. 1 (1996): 15-20; Mangione-Smith, R., Elliott, M.N., McDonald, L., and E.A. McGlynn. "An observational study of antibiotic prescribing behavior and the Hawthorne Effect." *Health Services Research,* Vol. 37, No. 6 (2002): 1603-23; Mujis, D. "Measuring teacher effectiveness: Some methodological reflections." *Educational Research and Evaluation*, Vol. 12, No. 1 (2006): 53–74; and McCarney, R., Warner, J., Iliffe, S., van Haselen, R., Griffin, M., and P. Fisher. "The Hawthorne Effect: a randomized, controlled trial." *BMC Medical Research Methodology*, Vol. 7, No. 30 (2007).

[137] For demonstrations that forensic examiners change their behavior when they know their performance is being monitored in particular ways, see Langenburg, G. "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." *Journal of Forensic Identification*, Vol. 59, No. 2 (2009).

orchestrate on a large scale.[138] On the other hand, test-blind proficiency tests have been used for DNA analysis,[139] and select labs have begun to implement this type of testing, in-house, as part of their quality assurance programs.[140] We note that test-blind proficiency testing is much easier to adopt in laboratories that have adopted "context management procedures" to reduce contextual bias.[141]

PCAST believes that test-blind proficiency testing of forensic examiners should be vigorously pursued, with the expectation that it should be in wide use, at least in large laboratories, within the next five years. However, PCAST believes that it is not yet realistic to require test-blind proficiency testing because the procedures for test-blind proficiency tests have not yet been designed and evaluated.

While only non-test-blind proficiency tests are used to support validity as applied, it is scientifically important to report this limitation, including to juries—because, as noted above, non-blind proficiency tests are likely to overestimate the accuracy because the examiners knew they were being tested.

4.7 Non-Empirical Views in the Forensic Community

While the scientific validity of metrological methods requires empirical demonstration of accuracy, there have historically been efforts in the forensic community to justify non-empirical approaches. This is of particular concern because such views are sometimes mistakenly codified in policies or practices. These heterodox views typically involve four recurrent themes, which we review below.

"Theories" of Identification

A common argument is that forensic practices should be regarded as valid because they rest on scientific "theories" akin to the fundamental laws of physics, that should be accepted because they have been tested and not "falsified."[142]

An example is the "Theory of Identification as it Relates to Toolmarks," issued in 2011 by the Association of Firearm and Tool Mark Examiners.[143,144] It states in its entirety:

[138] Some of the challenges associated with designing blind inter-laboratory proficiency tests may be addressed if the forensic laboratories were to move toward a system where an examiner's knowledge of a case were limited to domain-relevant information.

[139] See: Peterson, J.L., Lin, G., Ho, M., Chen, Y., and R.E. Gaensslen. "The feasibility of external blind DNA proficiency testing. II. Experience with actual blind tests." *Journal of Forensic Science,* Vol. 48, No. 1 (2003): 32-40.

[140] For example, the Houston Forensic Science Center has implemented routine, blind proficiency testing for its firearms examiners and chemistry analysis unit, and is planning to carry out similar testing for its DNA and latent print examiners.

[141] For background, see www.justice.gov/ncfs/file/888586/download.

[142] See: www.swggun.org/index.php?option=com_content&view=article&id=66:the-foundations-of-firearm-and-toolmark-identification&catid=13:other&Itemid=43 and www.justice.gov/ncfs/file/888586/download.

[143] Association of Firearm and Tool Mark Examiners. "Theory of Identification as it Relates to Tool Marks: Revised." *AFTE Journal*, Vol. 43, No. 4 (2011): 287.

[144] Firearms analysis is considered in detail in Chapter 5.

> *1. The theory of identification as it pertains to the comparison of toolmarks enables opinions of common origin to be made when the unique surface of two toolmarks are in "sufficient agreement."*
>
> *2. This "sufficient agreement" is related to the significant duplication of random toolmarks as evidenced by the correspondence of a pattern or combination of patterns of surface contours. Significance is determined by the comparative examination of two or more sets of surface contour patterns comprised of individual peaks, ridges and furrows. Specifically, the relative height or depth, width, curvature and spatial relationship of the individual peaks, ridges and furrows within one set of surface contours are defined and compare to the corresponding features in the second set of surface contours. Agreement is significant when the agreement in individual characteristics exceeds the best agreement demonstrated between toolmarks known to have been produced by different tools and is consistent with agreement demonstrated by toolmarks known to have been produced by the same tool. The statement that "sufficient agreement" exists between two toolmarks means that the agreement of individual characteristics is of a quantity and quality that the likelihood another tool could have made the mark is so remote as to be considered a practical impossibility.*
>
> *3. Currently the interpretation of individualization/identification is subjective in nature, founded on scientific principles and based on the examiner's training and experience.*

The statement is clearly not a scientific theory, which the National Academy of Sciences has defined as "a comprehensive explanation of some aspect of nature that is supported by a vast body of evidence."[145] Rather, it is a claim that examiners applying a subjective approach can accurately individualize the origin of a toolmark. Moreover, a "theory" is not what is needed. What is needed are empirical tests to see how well the method performs.

More importantly, the stated method is circular. It declares that an examiner may state that two toolmarks have a "common origin" when their features are in "sufficient agreement." It then defines "sufficient agreement" as occurring when the examiner considers it a "practical impossibility" that the toolmarks have different origins. (In response to PCAST's concern about this circularity, the FBI Laboratory replied that: "'Practical impossibility' is the certitude that exists when there is sufficient agreement in the quality and quantity of individual characteristics."[146] This answer did not resolve the circularity.)

Focus on 'Training and Experience' Rather Than Empirical Demonstration of Accuracy

Many practitioners hold an honest belief that they are able to make accurate judgments about identification based on their training and experience. This notion is explicit in the AFTE's *Theory of Identification*, which notes that interpretation is subjective in nature, "based on an examiner's training and experience." Similarly, the leading textbook on footwear analysis states,

> *Positive identifications may be made with as few as one random identifying characteristic, but only if that characteristic is confirmable; has sufficient definition, clarity, and features; is in the same location and*

[145] See: www.nas.edu/evolution/TheoryOrFact.html.
[146] Communication from FBI Laboratory to PCAST (June 6, 2016).

> *orientation on the shoe outsole; and <u>in the opinion of an experienced examiner, would not occur again on another shoe</u>.*[147] *[emphasis added]*

In effect, it says, positive identification depends on the examiner being *positive* about the identification.

"Experience" is an inadequate foundation for drawing judgments about whether two sets of features could have been produced by (or found on) different sources. Even if examiners could recall in sufficient detail all the patterns or sets of features that they have seen, they would have no way of knowing accurately in which cases two patterns actually came from different sources, because the correct answers are rarely known in casework.

The fallacy of relying on "experience" was evident in testimony by a former head of the FBI's fingerprint unit (discussed above) that the FBI had "an error rate of one per every 11 million cases," based on the fact that the agency was only aware of one mistake.[148] By contrast, recent empirical studies by the FBI Laboratory (discussed in Chapter 5) indicate error rates of roughly one in several hundred.

"Training" is an even weaker foundation. The mere fact that an individual has been trained in a method does not mean that the method itself is scientifically valid nor that the individual is capable of producing reliable answers when applying the method.

Focus on 'Uniqueness' Rather Than Accuracy

Many forensic feature-comparison disciplines are based on the premise that various sets of features (for example, fingerprints, toolmarks on bullets, human dentition, and so on) are "unique."[149]

[147] Bodziak, W. J. *Footwear Impression Evidence: Detection, Recovery, and Examination*. 2nd ed. CRC Press-Taylor & Francis, Boca Raton, Florida (2000).

[148] *U.S. v. Baines* 573 F.3d 979 (2009) at 984.

[149] For fingerprints, see, for example: Wertheim, Kasey. "Letter re: ACE-V: Is it scientifically reliable and accurate?" *Journal of Forensic Identification*, Vol. 52 (2002): 669 ("The law of biological uniqueness states that exact replication of any given organism cannot occur (nature never repeats itself), and, therefore, no biological entity will ever be exactly the same as another") and Budowle, B., Buscaglia, J., and R.S. Perlman. "Review of the scientific basis for friction ridge comparisons as a means of identification: committee findings and recommendations." *Forensic Science Communications*, Vol. 8 (2006) ("The use of friction ridge skin comparisons as a means of identification is based on the assumptions that the pattern of friction ridge skin is both unique and permanent"). For firearms, see, for example, Riva, F., and C. Christope. "Automatic comparison and evaluation of impressions left by a firearm on fired cartridge cases." *Journal of Forensic Sciences*, Vol. 59, (2014): 637 ("The ability to identify a firearm as the source of a questioned cartridge case or bullet is based on two tenets constituting the scientific foundation of the discipline. The first assumes the uniqueness of impressions left by the firearms") and SWGGUN Admissibility Resource Kit (ARK): Foundational Overview of Firearm/Toolmark Identification. available at: afte.org/resources/swggun-ark ("The basis for identification in Toolmark Identification is founded on the principle of uniqueness . . . wherein, all objects are unique to themselves and thus can be differentiated from one another"). For bitemarks, see, for example, Kieser, J.A., Bernal, V., Neil Waddell, J., and S. Raju. "The uniqueness of the human anterior dentition: a geometric morphometric analysis." *Journal of Forensic Sciences,* Vol. 52 (2007): 671-7 ("There are two postulates that underlie all bitemark analyses: first, that the characteristics of the anterior teeth involved in the bite are unique, and secondly, that this uniqueness is accurately recorded in the material bitten.") and Pretty, I.A. "Resolving Issues in Bitemark Analysis" in *Bitemark Evidence: A Color Atlas* R.B.J Dorian, Ed. CRC Press. Chicago (2011) ("Bitemark

The forensics science literature contains many "uniqueness" studies that go to great lengths to try to establish the correctness of this premise.[150] For example, a 2012 paper studied 39 Adidas Supernova Classic running shoes (size 12) worn by a single runner over 8 years, during which time he kept a running journal and ran over the same types of surfaces. [151] After applying black shoe polish to the soles of the shoes, the author asked the runner to carefully produce tread marks on sheets of legal paper on a hardwood floor. The author showed that it was possible to identify small identifying differences between the tread marks produced by different pairs of shoes.

Yet, uniqueness studies miss the fundamental point. The issue is not whether *objects* or *features* differ; they surely do if one looks at a fine enough level. The issue is how well and under what circumstances *examiners* applying a given metrological method can reliably *detect* relevant differences in features to reliably identify whether they share a common source. Uniqueness studies, which focus on the properties of features themselves, can therefore never establish whether a particular *method* for measuring and comparing features is foundationally valid. Only empirical studies can do so.

Moreover, it is not *necessary* for features to be unique in order for them to be useful in narrowing down the source of a feature. Rather, it is essential that there be empirical evidence about how often a method incorrectly attributes the source of a feature.

Decoupling Conclusions about Identification from Estimates of Accuracy

Finally, some hold the view that, when the application of a scientific method leads to a conclusion of an association or proposed identification, it is *unnecessary* to report in court the reliability of the method.[152] As a rationale, it is sometimes argued that it is impossible to measure error rates perfectly or that it is impossible to know the error rate in the *specific* case at hand.

This notion is contrary to the fundamental principle of scientific validity in metrology—namely, that the claim that two objects have been compared and found to have the same property (length, weight, or fingerprint pattern) is meaningless without quantitative information about the reliability of the comparison process.

It is standard practice to study and report error rates in medicine—both to establish the reliability of a method in principle and to assess its implementation in practice. No one argues that measuring or reporting clinical error rates is inappropriate because they might not perfectly reflect the situation for a *specific* patient. If

analysis is based on two postulates: (a) the dental characteristics of anterior teeth involved in biting are unique among individuals, and (b) this asserted uniqueness is transferred and recorded in the injury.").

[150] Some authors have criticized attempts to affirm the uniqueness proposition based on observations, noting that they rest on pure inductive reasoning, a method for scientific investigation that "fell out of favour during the epoch of Sir Francis Bacon in the 16th century." Page, M., Taylor, J., and M. Blenkin. "Uniqueness in the forensic identification sciences—fact or fiction?" *Forensic Science International*, Vol. 206 (2011): 12-8.

[151] Wilson, H.D. "Comparison of the individual characteristics in the outsoles of thirty-nine pairs of Adidas Supernova Classic shoes." *Journal of Forensic Identification*, Vol. 62, No. 3 (2012): 194-204.

[152] See: www.justice.gov/olp/file/861936/download.

transparency about error rates is appropriate for matching blood types before a transfusion, it is appropriate for matching forensic samples—where errors may have similar life-threatening consequences.

We return to this topic in Chapter 8, where we observe that the DOJ's recent proposed guidelines on expert testimony are based, in part, on this scientifically inappropriate view.

4.8 Empirical Views in the Forensic Community

Although some in the forensic community continue to hold views such as those described in the previous section, a growing segment of the forensic science community has responded to the 2009 NRC report with an increased recognition of the need for empirical studies and with initial efforts to undertake them. Examples include published research studies by forensic scientists, assessments of research needs by Scientific Working Groups and OSAC committees, and statements from the NCFS.

Below we highlight several examples from recent papers by forensic scientists:

- *Researchers at the National Academy of Sciences and elsewhere (e.g., Saks & Koehler, 2005; Spinney, 2010) have argued that there is an urgent need to develop objective measures of accuracy in fingerprint identification. Here we present such data.*[153]

- *Tool mark impression evidence, for example, has been successfully used in courts for decades, but its examination has lacked scientific, statistical proof that would independently corroborate conclusions based on morphology characteristics (2–7). In our study, we will apply methods of statistical pattern recognition (i.e., machine learning) to the analysis of toolmark impressions.*[154]

- *The NAS report calls for further research in the area of bitemarks to demonstrate that there is a level of probative value and possibly restricting the use of analyses to the exclusion of individuals. This call to respond must be heard if bite-mark evidence is to be defensible as we move forward as a discipline.*[155]

- *The National Research Council of the National Academies and the legal and forensic sciences communities have called for research to measure the accuracy and reliability of latent print examiners' decisions, a challenging and complex problem in need of systematic analysis. Our research is focused on the development of empirical approaches to studying this problem.*[156]

[153] Tangen, J.M., Thompson, M.B., and D.J. McCarthy. "Identifying fingerprint expertise." *Psychological Science*, Vol. 22, No. 8 (2011): 995-7.

[154] Petraco, N.D., Shenkin, P., Speir, J., Diaczuk, P., Pizzola, P.A., Gambino, C., and N. Petraco. "Addressing the National Academy of Sciences' Challenge: A Method for Statistical Pattern Comparison of Striated Tool Marks." *Journal of Forensic Sciences*, Vol. 57 (2012): 900-11.

[155] Pretty, I.A., and D. Sweet. "A paradigm shift in the analysis of bitemarks." *Forensic Science International*, Vol. 201 (2010): 38-44.

[156] Ulery, B.T., Hicklin, R.A., Buscaglia, J., and M.A., Roberts. "Accuracy and reliability of forensic latent fingerprint decisions." *PNAS*, Vol. 108, No. 19 (2011): 7733-8.

- *We believe this report should encourage the legal community to require that the emerging field of forensic neuroimaging, including fMRI based lie detection, have a proper scientific foundation before being admitted in courts.*[157]

- *An empirical solution which treats the system [referring to voiceprints] as a black box and its output as point values is therefore preferred.*[158]

Similarly, the OSAC and other groups have acknowledged critical research gaps in the evidence supporting various forensic science disciplines and have begun to develop plans to close some of these gaps. We highlight several examples below:

- *While validation studies of firearms and toolmark analysis schemes have been conducted, most have been relatively small data sets. If a large study were well designed and has sufficient participation, it is our anticipation that similar lessons could be learned for the firearms and toolmark discipline.*[159]

- *We are unaware of any study that assesses the overall firearm and toolmark discipline's ability to correctly/consistently categorize evidence by class characteristics, identify subclass marks, and eliminate items using individual characteristics.*[160]

- *Currently there is not a reliable assessment of the discriminating strength of specific friction ridge feature types.*[161]

- *To date there is little scientific data that quantifies the overall risk of close non-matches in AFIS databases. It is difficult to create standards regarding sufficiency for examination or AFIS search searching without this type of research.*[162]

[157] Langleben, D.D., and J.C. Moriarty. "Using brain imaging for lie detection: Where science, law, and policy collide." *Psychology, Public Policy, and Law*, Vol. 19, No. 2 (2013): 222–34.

[158] Morrison, G.S., Zhang, C., and P. Rose. "An empirical estimate of the precision of likelihood ratios from a forensic-voice-comparison system." *Forensic Science International*, Vol. 208, (2011): 59–65.

[159] OSAC Research Needs Assessment Form. "Study to Assess The Accuracy and Reliability of Firearm and Toolmark." Issued October 2015 (Approved January 2016). Available at: www.nist.gov/forensics/osac/upload/FATM-Research-Needs-Assessment_Blackbox.pdf.

[160] OSAC Research Needs Assessment Form. "Assessment of Examiners' Toolmark Categorization Accuracy." Issued October 2015 (Approved January 2016). Available at: www.nist.gov/forensics/osac/upload/FATM-Research-Needs-Assessment_Class-and-individual-marks.pdf.

[161] OSAC Research Needs Assessment Form. "Assessing the Sufficiency and Strength of Friction Ridge Features." Issued October 2015. Available at: www.nist.gov/forensics/osac/upload/FRS-Research-Need-Assessment-of-Features.pdf.

[162] OSAC Research Needs Assessment Form. "Close Non-Match Assessment." Issued October 2015. Available at: www.nist.gov/forensics/osac/upload/FRS-Research-Need-Close-Non-Match-Assessment.pdf.

- *Research is needed that studies whether sequential unmasking reduces the negative effects of bias during latent print examination.*[163]

- *The IAI has, for many years, sought support for research that would scientifically validate many of the comparative analyses conducted by its member practitioners. While there is a great deal of empirical evidence to support these exams, independent validation has been lacking.*[164]

The National Commission on Forensic Science has similarly recognized the need for rigorous empirical evaluation of forensic methods in a Views Document approved by the commission:

> *All forensic science methodologies should be evaluated by an independent scientific body to characterize their capabilities and limitations in order to accurately and reliably answer a specific and clearly defined forensic question.*[165]

PCAST applauds this growing focus on empirical evidence. We note that increased research funding will be needed to achieve these critical goals (see Chapter 6).

4.9 Summary of Scientific Findings

We summarize our scientific findings concerning the scientific criteria for foundational validity and validity as applied.

Finding 1: Scientific Criteria for Scientific Validity of a Forensic Feature-Comparison Method

(1) Foundational validity. To establish foundational validity for a forensic feature-comparison method, the following elements are required:

(a) a reproducible and consistent procedure for (i) identifying features in evidence samples; (ii) comparing the features in two samples; and (iii) determining, based on the similarity between the features in two sets of features, whether the samples should be declared to be likely to come from the same source ("matching rule"); and

(b) empirical estimates, from appropriately designed studies from multiple groups, that establish (i) the method's false positive rate—that is, the probability it declares a proposed identification between samples that actually come from different sources and (ii) the method's sensitivity—that is, the probability it declares a proposed identification between samples that actually come from the same source.

[163] OSAC Research Needs Assessment Form. "ACE-V Bias." Issued October 2015. Available at: www.nist.gov/forensics/osac/upload/FRS-Research-Need-ACE-V-Bias.pdf.

[164] International Association for Identification. Letter to Patrick J. Leahy, Chairman, Senate Committee on the Judiciary, March 18, 2009. Available at: www.theiai.org/current_affairs/nas_response_leahy_20090318.pdf.

[165] National Commission on Forensic Science: "Views of the Commission Technical Merit Evaluation of Forensic Science Methods and Practices." Available at: www.justice.gov/ncfs/file/881796/download.

As described in Box 4, scientific validation studies should satisfy a number of criteria: (a) they should be based on sufficiently large collections of known and representative samples from relevant populations; (b) they should be conducted so that the examinees have no information about the correct answer; (c) the study design and analysis plan should be specified in advance and not modified afterwards based on the results; (d) the study should be conducted or overseen by individuals or organizations with no stake in the outcome; (e) data, software and results should be available to allow other scientists to review the conclusions; and (f) to ensure that the results are robust and reproducible, there should be multiple independent studies by separate groups reaching similar conclusions.

Once a method has been established as foundationally valid based on adequate empirical studies, claims about the method's accuracy and the probative value of proposed identifications, in order to be valid, must be based on such empirical studies.

For objective methods, foundational validity can be established by demonstrating the reliability of each of the individual steps (feature identification, feature comparison, matching rule, false match probability, and sensitivity).

For subjective methods, foundational validity can be established *only* through black-box studies that measure how often many examiners reach accurate conclusions across many feature-comparison problems involving samples representative of the intended use. In the absence of such studies, a subjective feature-comparison method cannot be considered scientifically valid.

Foundational validity is a *sine qua non*, which can only be shown through empirical studies. Importantly, good professional practices—such as the existence of professional societies, certification programs, accreditation programs, peer-reviewed articles, standardized protocols, proficiency testing, and codes of ethics—cannot substitute for empirical evidence of scientific validity and reliability.

(2) Validity as applied. Once a forensic feature-comparison method has been established as foundationally valid, it is necessary to establish its validity as applied in a given case.

As described in Box 5, validity as applied requires that: (a) the forensic examiner must have been shown to be *capable* of reliably applying the method, as shown by appropriate proficiency testing (see Section 4.6), and must *actually* have done so, as demonstrated by the procedures actually used in the case, the results obtained, and the laboratory notes, which should be made available for scientific review by others; and (b) assertions about the probative value of proposed identifications must be scientifically valid—including that examiners should report the overall false positive rate and sensitivity for the method established in the studies of foundational validity; demonstrate that the samples used in the foundational studies are relevant to the facts of the case; where applicable, report probative value of the observed match based on the specific features observed in the case; and not make claims or implications that go beyond the empirical evidence.

5. Evaluation of Scientific Validity for Seven Feature-Comparison Methods

In the previous chapter, we described the scientific criteria that a forensic feature-comparison method must meet to be considered scientifically valid and reliable, and we underscored the need for empirical evidence of accuracy and reliability.

In this chapter, we illustrate the meaning of these criteria by applying them to six specific forensic feature-comparison methods: (1) DNA analysis of single-source and simple-mixture samples, (2) DNA analysis of complex-mixture samples, (3) bitemarks, (4) latent fingerprints, (5) firearms identification, and (6) footwear analysis.[166] For a seventh forensic feature- comparison method, hair analysis, we do not undertake a full evaluation, but review a recent evaluation by the DOJ.

We evaluate whether these methods have been established to be foundationally valid and reliable and, if so, what estimates of accuracy should accompany testimony concerning a proposed identification, based on current scientific studies. We also briefly discuss some issues related to validity as applied.

PCAST compiled a list of 2019 papers from various sources—including bibliographies prepared by the National Science and Technology Council's Subcommittee on Forensic Science, the relevant Scientific Working Groups (predecessors to the current OSAC),[167] and the relevant OSAC committees; submissions in response to PCAST's request for information from the forensic-science stakeholder community; and our own literature searches.[168] PCAST members and staff identified and reviewed those papers that were relevant to establishing scientific validity. After reaching a set of initial conclusions, input was obtained from the FBI Laboratory and individual scientists at NIST, as well as other experts—including asking them to identify additional papers supporting scientific validity that we might have missed.

For each of the methods, we provide a brief overview of the methodology, discuss background information and studies, and review evidence for scientific validity.

As discussed in Chapter 4, objective methods have well-defined procedures to (1) identify the features in samples, (2) measure the features, (3) determine whether the features in two samples match to within a stated measurement tolerance (matching rule), and (4) estimate the probability that samples from different sources would match (false match probability). It is possible to examine each of these separate steps for their validity

[166] The American Association for the Advancement of Science (AAAS) is conducting an analysis of the underlying scientific bases for the forensic tools and methods currently used in the criminal justice system. As of September 1, 2016 no reports have been issued. See: www.aaas.org/page/forensic-science-assessments-quality-and-gap-analysis.
[167] See: www.nist.gov/forensics/workgroups.cfm.
[168] See: www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensics_references.pdf.

and reliability. Of the six methods considered in this chapter, only the first two methods (involving DNA analysis) employ objective methods. The remaining four methods are subjective.

For subjective methods, the procedures are not precisely defined, but rather involve substantial expert human judgment. Examiners may focus on certain features while ignoring others, may compare them in different ways, and may have different standards for declaring proposed identification between samples. As described in Chapter 4, the sole way to establish foundational validity is through multiple independent "black-box" studies that measure how often examiners reach accurate conclusions across many feature-comparison problems involving samples representative of the intended use. In the absence of such studies, a feature-comparison method cannot be considered scientifically valid.

PCAST found few black-box studies appropriately designed to assess scientific validity of subjective methods. Two notable exceptions, discussed in this chapter, were a study on latent fingerprints conducted by the FBI Laboratory and a study on firearms identification sponsored by the Department of Defense and conducted by the Department of Energy's Ames Laboratory.

We considered whether proficiency testing, which is conducted by commercial organizations for some disciplines, could be used to establish foundational validity. We concluded that it could not, at present, for several reasons. First, proficiency tests are not intended to establish foundational validity. Second, the test problems or test sets used in commercial proficiency tests are not at present routinely made public—making it impossible to ascertain whether the tests appropriately assess the method across the range of applications for which it is used. The publication and critical review of methods and data is an essential component in establishing scientific validity. Third, the dominant company in the market, Collaborative Testing Services, Inc. (CTS), explicitly states that its proficiency tests are not appropriate for estimating error rates of a discipline, because (a) the test results, which are open to anyone, may not reflect the skills of forensic practitioners and (b) "the reported results do not reflect 'correct' or 'incorrect' answers, but rather responses that agree or disagree with the consensus conclusions of the participant population."[169] Fourth, the tests for forensic feature-comparison methods typically consist of only one or two problems each year. Fifth, "easy tests are favored by the community," with the result that tests that are too challenging could jeopardize repeat business for a commercial vendor.[170]

[169] See: www.ctsforensics.com/assets/news/CTSErrorRateStatement.pdf.

[170] PCAST thanks Collaborative Testing Services, Inc. (CTS) President Christopher Czyryca for helpful conversations concerning proficiency testing. Czyryca explained that that (1) CTS defines consensus as at least 80 percent agreement among respondents and (2) proficiency testing for latent fingerprints only occasionally involves a problem in which a questioned print matches *none* of the possible answers. Czyryca noted that the forensic community disfavors more challenging tests—and that testing companies are concerned that they could lose business if their tests are viewed as too challenging. An example of a "challenging" test is the very important scenario in which *none* of the questioned samples match any of the known samples: because examiners may expect they should find *some* matches, such scenarios provide an opportunity to assess how often examiners declare false-positive matches. (See also presentation to the National Commission on Forensic Science by CTS President Czyryca, noting that "Easy tests are favored by the community." www.justice.gov/ncfs/file/761061/download.)

PCAST's observations and findings below are largely consistent with the conclusions of earlier NRC reports.[171]

5.1 DNA Analysis of Single-source and Simple-mixture samples

DNA analysis of single-source and simple mixture samples includes excellent examples of objective methods whose foundational validity has been properly established.[172]

Methodology

DNA analysis involves comparing DNA profiles from different samples to see if a known sample may have been the source of an evidentiary sample.

To generate a DNA profile, DNA is first chemically *extracted* from a sample containing biological material, such as blood, semen, hair, or skin cells. Next, a predetermined set of DNA segments ("loci") containing small repeated sequences[173] are *amplified* using the Polymerase Chain Reaction (PCR), an enzymatic process that replicates a targeted DNA segment over and over to yield millions of copies. After amplification, the lengths of the resulting DNA fragments are *measured* using a technique called capillary electrophoresis, which is based on the fact that longer fragments move more slowly than shorter fragments through a polymer solution. The raw data collected from this process are analyzed by a software program to produce a graphical image (an electropherogram) and a list of numbers (the DNA profile) corresponding to the sizes of the each of fragments (by comparing them to known "molecular size standards").

As currently practiced, the method uses 13 specific loci and the amplification process is designed so that the DNA fragments corresponding to different loci occupy different size ranges—making it simple to recognize which fragments come from each locus.[174] At each locus, every human carries two variants (called "alleles")—one inherited from his or her mother, one from his or her father—that may be of different lengths or the same length.[175]

[171] National Research Council. *Strengthening Forensic Science in the United States: A Path Forward.* The National Academies Press. Washington DC. (2009). National Research Council, *Ballistic Imaging*. The National Academies Press. Washington DC. (2008).

[172] Forensic DNA analysis belongs to two parent disciplines—metrology and human molecular genetics—and has benefited from the extensive application of DNA technology in biomedical research and medical application.

[173] The repeats, called short tandem repeats (STRs), consist of consecutive repeated copies of a segments of 2-6 base pairs.

[174] The current kit used by the FBI (Identifiler Plus) has 16 total loci: 15 STR loci and the amelogenin locus. A kit that will be implemented later this year has 24 loci.

[175] The FBI announced in 2015 that it plans to expand the core loci by adding seven additional loci commonly used in databases in other countries. (Population data have been published for the expanded set, including frequencies in 11 ethnic populations www.fbi.gov/about-us/lab/biometric-analysis/codis/expanded-fbi-str-2015-final-6-16-15.pdf.) Starting in 2017, these loci will be required for uploading and searching DNA profiles in the national system. The expanded data in each profile are expected to provide greater discrimination potential for identification, especially in matching samples with only partial DNA profiles, missing person inquiries, and international law enforcement and counterterrorism cases.

Analysis of single-source samples

DNA analysis of a sample from a single individual is an objective method. In addition to the laboratory protocols being precisely defined, the interpretation also involves little or no human judgment.

An examiner can assess if a sample came from a single source based on whether the DNA profile typically contains, for each locus, exactly one fragment from each chromosome containing the locus—which yields one or two distinct fragment lengths from each locus.[176] The DNA profile can then be compared with the DNA profile of a known suspect. It can also be entered into the FBI's National DNA Index System (NDIS) and searched against a database of DNA profiles from convicted offenders (and arrestees in more than half of the states) or unsolved crimes.

Two DNA profiles are declared to match if the lists of alleles are the same.[177] The probability that two DNA profiles from *different* sources would have the same DNA profile (the random match probability) is then calculated based on the empirically measured frequency of each allele and established principles of population genetics (see p. 53).[178]

Analysis of simple mixtures

Many sexual assault cases involve DNA mixtures of two individuals, where one individual (i.e., the victim) is known. DNA analysis of these simple mixtures is also relatively straightforward. Methods have been used for 30 years to differentially extract DNA from sperm cells vs. vaginal epithelial cells, making it possible to generate DNA profiles from the two sources. Where the two cell types are the same but one contributor is known, the alleles of the known individual can be subtracted from the set of alleles identified in the mixture.[179]

Once the known source is removed, the analysis of the unknown sample then proceeds as above for single-source samples. Like the analysis of single-source samples, the analysis of simple mixtures is a largely objective method.

[176] The examiner reviews the electropherogram to determine whether each of the peaks is a true allelic peak or an artifact (e.g., background noise in the form of stutter, spikes, and other phenomena) and to determine whether more than one individual could have contributed to the profile. In rare cases, an individual may have two fragments at a locus due to rare copy-number variation in the human genome.

[177] When only a partial profile could be generated from the evidence sample (for example, in cases with limited quantities of DNA, degradation of the sample, or the presence of PCR inhibitors), an examiner may also report an "inclusion" if the partial profile is *consistent* with the DNA profile obtained from a reference sample. An examiner may also report an inclusion when the DNA results from a reference sample are present in a mixture. These cases generally require significantly more human analysis and interpretation than single-source samples.

[178] Random match probabilities can also be expressed in terms of a likelihood ratio (LR), which is the ratio of (1) the probability of observing the DNA profile if the individual in question is the source of the DNA sample and (2) the probability of observing the DNA profile if the individual in question is *not* the source of the DNA sample. In the situation of a single-source sample, the LR should be simply the reciprocal of the random match probability (because the first probability in the LR is 1 and the second probability is the random match probability).

[179] In many cases, DNA will be present in the mixture in sufficiently different quantities so that the peak heights in the electropherogram from the two sources will be distinct, allowing the examiner to more readily separate out the sources.

Foundational Validity

To evaluate the foundational validity of an objective method (such as single-source and simple mixture analysis), one can examine the reliability of each of the individual steps rather than having to rely on black-box studies.

Single-source samples

Each step in the analysis is objective and involves little or no human judgment.

(1) *Feature identification*. In contrast to the other methods discussed in this report, the features used in DNA analysis (the fragments lengths of the loci) are defined *in advance*.

(2) *Feature measurement and comparison*. PCR amplification, invented in 1983, is widely used by tens of thousands of molecular biology laboratories, including for many medical applications in which it has been rigorously validated. Multiplex PCR kits designed by commercial vendors for use by forensic laboratories must be validated both externally (through developmental validation studies published in peer reviewed publication) and internally (by each lab that wishes to use the kit) before they may be used.[180] Fragment sizes are measured by an automated procedure whose variability is well characterized and small; the standard deviation is approximately 0.05 base pairs, which provides highly reliable measurements.[181,182] Developmental validation studies were performed—including by the FBI—to verify the accuracy, precision, and reproducibility of the procedure.[183,184]

[180] Laboratories that conduct forensic DNA analysis are required to follow FBI's Quality Assurance Standards for DNA Testing Laboratories as a condition of participating in the National DNA Index System (www.fbi.gov/about-us/lab/biometric-analysis/codis/qas-standards-for-forensic-dna-testing-laboratories-effective-9-1-2011). FBI's Scientific Working Group on DNA Analysis Methods (SWGDAM) has published guidelines for laboratories in validating procedures consistent the FBI's Quality Assurance Standards (QAS). SWGDAM Validation Guidelines for DNA Analysis Methods, December 2012. See: media.wix.com/ugd/4344b0_cbc27d16dcb64fd88cb36ab2a2a25e4c.pdf.

[181] Forensic laboratories typically use genetic analyzer systems developed by the Applied Biosystems group of Thermo-Fisher Scientific (ABI 310, 3130, or 3500).

[182] To incorrectly estimate a fragment length by 1 base pair (the minimum size difference) requires a measurement error of 0.5 base pair, which corresponds to 10 standard deviations. Moreover, alleles typically differ by at least 4 base pairs (although some STR loci have fairly common alleles that differ by 1 or 2 nucleotides).

[183] For examples of these studies see: Budowle, B., Moretti, T.R., Keys, K.M., Koons, B.W., and J.B. Smerick. "Validation studies of the CTT STR multiplex system." *Journal of Forensic Sciences,* Vol. 42, No. 4 (1997): 701-7; Kimpton, C.P., Oldroyd, N.J., Watson, S.K., Frazier, R.R., Johnson, P.E., Millican, E.S., Urguhart, A., Sparkes, B.L., and P. Gill. "Validation of highly discriminating multiplex short tandem repeat amplification systems for individual identification." *Electrophoresis*, Vol. 17, No. 8 (1996): 1283-93; Lygo, J.E., Johnson, P.E., Holdaway, D.J., Woodroffe, S., Whitaker, J.P., Clayton, T.M., Kimpton, C.P., and P. Gill. "The validation of short tandem repeat (STR) loci for use in forensic casework." *International Journal of Legal Medicine,* Vol. 107, No. 2 (1994): 77-89; and Fregeau, C.J., Bowen, K.L., and R.M. Fourney. "Validation of highly polymorphic fluorescent multiplex short tandem repeat systems using two generations of DNA sequencers." *Journal of Forensic Sciences*, Vol. 44, No. 1 (1999): 133-66.

[184] For example, a 2001 study that compared the performance characteristics of several commercially available STR testing kits tested the consistency and reproducibility of results using previously typed case samples, environmentally insulted samples, and body fluid samples deposited on various substrates. The study found that all of the kits could be used to amplify and type STR loci successfully and that the procedures used for each of the kits were robust and valid. No evidence

(3) *Feature comparison*. For single-source samples, there are clear and well-specified "matching rules" for declaring whether the DNA profiles match. When complete DNA profiles are searched against the NDIS at "high stringency," a "match" is returned only when each allele in the unknown profile is found to match an allele of the known profile, and *vice versa*. When partial DNA profiles obtained from a partially degraded or contaminated sample are searched at "moderate stringency," candidate profiles are returned if each of the alleles in the unknown profile is found to match an allele of the known profile.[185,186]

(4) *Estimation of random match probability*. The process for calculating the random match probability (that is, the probability of a match occurring by chance) is based on well-established principles of population genetics and statistics. The frequencies of the individual alleles were obtained by the FBI based on DNA profiles from approximately 200 unrelated individuals from each of six population groups and were evaluated prior to use.[187] The frequency of an overall pattern of alleles—that is, the random match probability—is typically estimated by multiplying the frequencies of the individual loci, under the assumption that the alleles are independent of one another.[188] The resulting probability is typically less than 1 in 10 billion, excluding the possibility of close relatives.[189] (Note: Multiplying the frequency of alleles can overstates the rarity of a pattern because the alleles are not completely independent, owing

of false positive or false negative results and no substantial evidence of preferential amplification within a locus were found for any of the testing kits. Moretti, T.R., Baumstark, A.L., Defenbaugh, D.A., Keys, K.M., Smerick, J.B., and B. Budowle. "Validation of Short Tandem Repeats (STRs) for forensic usage: performance testing of fluorescent multiplex STR systems and analysis of authentic and simulated forensic samples." *Journal of Forensic Sciences*, Vol. 46, No. 3 (2001): 647-60.

[185] See: FBI's Frequently Asked Questions (FAQs) on the CODIS Program and the National DNA Index System. www.fbi.gov/about-us/lab/biometric-analysis/codis/codis-and-ndis-fact-sheet.

[186] Contaminated samples are not retained in NDIS.

[187] The initial population data generated by FBI included data for 6 ethnic populations with database sizes of 200 individuals. See: Budowle, B., Moretti, T.R., Baumstark, A.L., Defenbaugh, D.A., and K.M. Keys. "Population data on the thirteen CODIS core short tandem repeat loci in African Americans, U.S. Caucasians, Hispanics, Bahamians, Jamaicans, and Trinidadians." *Journal of Forensic Sciences*, Vol. 44, No. 6 (1999): 1277-86 and Budowle, B., Shea, B., Niezgoda, S., and R. Chakraborty. "CODIS STR loci data from 41 sample populations." *Journal of Forensic Sciences,* Vol. 46, No. 3 (2001): 453-89. Errors in the original database were reported in July 2015 (Erratum, *Journal of Forensic Sciences*, Vol. 60, No. 4 (2015): 1114-6, the impact of these discrepancies on profile probability calculations were assessed (and found to be less than a factor of 2 in a full profile), and the allele frequency estimates were amended accordingly. At the same time as amending the original datasets, the FBI Laboratory also published expanded datasets in which the original samples were retyped for additional loci. In addition, the population samples that were originally studied at other laboratories were typed for additional loci, so the full dataset includes 9 populations. These "expanded" datasets are in use at the FBI Laboratory and can be found at www.fbi.gov/about-us/lab/biometric-analysis/codis/expanded-fbi-str-final-6-16-15.pdf.

[188] More precisely, the frequency at each locus is calculated first. If the locus has two copies of the same allele with frequency p, the frequency is calculated as p^2. If the locus has two different alleles with respective frequencies p and q, the frequency is calculated as 2pq. The frequency of the overall pattern is calculated by multiplying together the values for the individual loci.

[189] The random match probability will be higher for close relatives. For identical twins, the DNA profiles are expected to match perfectly. For first degree relatives, the random match probability may be on the order of 1 in 100,000 when examining the 13 CODIS core STR loci. See: Butler, J.M. "The future of forensic DNA analysis." *Philosophical Transactions of the Royal Society B,* 370: 20140252 (2015).

to population substructure. A 1996 NRC report concluded that the effect of population substructure on the calculated value was likely to be within a factor of 10 (for example, for a random match probability estimate of 1 in 10 million, the true probability is highly likely to be between 1 in 1 million and 1 in 100 million).[190] However, a recent study by NIST scientists suggests that the variation may be substantially greater than 10-fold.[191] The random match probability should be calculated using an appropriate statistical formula that takes account of population substructure.[192])

Simple mixtures

The steps for analyzing simple mixtures are the same as for analyzing single-source samples, up until the point of interpretation. DNA profiles that contain a mixture of two contributors, where one contributor is known, can be interpreted in much the same way as single-source samples. This occurs frequently in sexual assault cases, where a DNA profile contains a mixture of DNA from the victim and the perpetrator. Methods that are used to differentially extract DNA from sperm cells vs. vaginal epithelial cells in sexual assault cases are well-established.[193] Where the two cell types are the same, one DNA source may be dominant, resulting in a distinct contrast in peak heights between the two contributors; in these cases, the alleles from both the major contributor (corresponding to the larger allelic peaks) and the minor contributor can usually be reliably interpreted, provided the proportion of the minor contributor is not too low.[194]

Validity as Applied

While DNA analysis of single-source samples and simple mixtures is a foundationally valid and reliable method, it is not infallible in practice. Errors can and do occur in DNA testing. Although the probability that two samples from different sources have the same DNA profile is tiny, the chance of human error is much higher. Such errors may stem from sample mix-ups, contamination, incorrect interpretation, and errors in reporting.[195]

[190] National Research Council. *The Evaluation of Forensic DNA Evidence.* The National Academies Press. Washington DC. (1996). Goode, M. "Some observations on evidence of DNA frequency." *Adelaide Law Review,* Vol. 23 (2002): 45-77.

[191] Gittelson, S. and J. Buckleton. "Is the factor of 10 still applicable today?" Presentation at the 68th Annual American Academy of Forensic Sciences Scientific Meeting, 2016. See: www.cstl.nist.gov/strbase/pub_pres/Gittelson-AAFS2016-Factor-of-10.pdf.

[192] Balding, D.J., and R.A. Nichols. "DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands." *Forensic Science International*, Vol. 64 (1994): 125-140.

[193] Gill, P., Jeffreys, A.J., and D.J. Werrett. "Forensic application of DNA 'fingerprints.'" *Nature*, Vol. 318, No. 6046 (1985): 577-9.

[194] Clayton, T.M., Whitaker, J.P., Sparkes, R., and P. Gill. "Analysis and interpretation of mixed forensic stains using DNA STR profiling." *Forensic Science International*, Vol. 91, No. 1 (1998): 55-70.

[195] Krimsky, S., and T. Simoncelli. *Genetic Justice: DNA Data Banks, Criminal Investigations, and Civil Liberties.* Columbia University Press, (2011). Perhaps the most spectacular human error to date involved the German government's investigation of the "Phantom of Heilbronn," a woman whose DNA appeared at the scenes of more than 40 crimes in three countries, including 6 murders, several muggings and dozens of break-ins over the course of more than a decade. After an effort that included analyzing DNA samples from more than 3,000 women from four countries and that cost $18 million, authorities discovered that the woman of interest was a worker in the Austrian factory that fabricated the swabs used in DNA collection. The woman had inadvertently contaminated a large number of swabs with her own DNA, which was thus found in many DNA tests.

To minimize human error, the FBI requires, as a condition of participating in NDIS, that laboratories follow the FBI's Quality Assurance Standards (QAS).[196] Before the results of the DNA analysis can be compared, the examiner is required to run a series of controls to check for possible contamination and ensure that the PCR process ran properly. The QAS also requires semi-annual proficiency testing of all DNA analysts that perform DNA testing for criminal cases. The results of the tests do not have to be published, but the laboratory must retain the results of the tests, any discrepancies or errors made, and corrective actions taken.[197]

Forensic practitioners in the U.S. do not typically report quality issues that arise in forensic DNA analysis. By contrast, error rates in medical DNA testing are commonly measured and reported.[198] Refreshingly, a 2014 paper from the Netherlands Forensic Institute (NFI), a government agency, reported a comprehensive analysis of all "quality issue notifications" encountered in casework, categorized by type, source and impact.[199,200] The authors call for greater "transparency" and "culture change," writing that:

> *Forensic DNA casework is conducted worldwide in a large number of laboratories, both private companies and in institutes owned by the government. Quality procedures are in place in all laboratories, but the nature of the quality system varies a lot between the different labs. In particular, there are many forensic DNA laboratories that operate without a quality issue notification system like the one described in this paper. In our experience, such a system is extremely important for the detection and proper handling of errors. This is crucial in forensic casework that can have a major impact on people's lives. We therefore propose that the implementation of a quality issue notification system is necessary for any laboratory that is involved in forensic DNA casework.*
>
> *Such system can only work in an optimal way, however, when there is a blame-free culture in the laboratory that extends to the police and the legal justice system. People have a natural tendency to hide their mistakes, and it is essential to create an atmosphere where there are no adverse personal consequences when mistakes are reported. The management should take the lead in this culture change...*
>
> *As far as we know, the NFI is the first forensic DNA laboratory in the world to reveal such detailed data and reports. It shows that this is possible without any disasters or abuse happening, and there are no*

[196] FBI. "Quality assurance standards for forensic DNA testing laboratories." (2011). See: www.fbi.gov/about-us/lab/biometric-analysis/codis/qas-standards-for-forensic-dna-testing-laboratories-effective-9-1-2011.

[197] Ibid., Sections 12, 13, and 14.

[198] See, for example: Plebani, M., and P. Carroro. "Mistakes in a stat laboratory: types and frequency." *Clinical Chemistry,* Vol. 43 (1997): 1348-51; Stahl, M., Lund, E.D., and I. Brandslund. "Reasons for a laboratory's inability to report results for requested analytical tests." *Clinical Chemistry,* Vol. 44 (1998): 2195-7; Hofgartner, W.T., and J.F. Tait. "Frequency of problems during clinical molecular-genetic testing." *American Journal of Clinical Pathology,* Vol. 112 (1999): 14-21; and Carroro, P., and M. Plebani. "Errors in a stat laboratory: types and frequencies 10 years later." *Clinical Chemistry,* Vol. 53 (2007): 1338-42.

[199] Kloosterman, A., Sjerps, M., and A. Quak. "Error rates in forensic DNA analysis: Definition, numbers, impact and communication." *Forensic Science International: Genetics*, Vol. 12 (2014): 77-85 and J.M. Butler "DNA Error Rates" presentation at the International Forensics Symposium, Washington, D.C. (2015). www.cstl.nist.gov/strbase/pub_pres/Butler-ErrorManagement-DNA-Error.pdf.

[200] The Netherlands uses an "inquisitorial" approach to method of criminal justice rather than the adversarial system used in the U.S. Concerns about having to explain quality issues in court may explain in part why U.S. laboratories do not routinely report quality issues.

> *reasons for nondisclosure. As mentioned in the introduction, in laboratory medicine publication of data on error rates has become standard practice. Quality failure rates in this domain are comparable to ours.*

Finally, we note that there is a need to improve proficiency testing. There are currently no requirements concerning how challenging the proficiency tests should be. The tests should be representative of the full range of situations likely to be encountered in casework.

Finding 2: DNA Analysis

Foundational validity. PCAST finds that DNA analysis of single-source samples or simple mixtures of two individuals, such as from many rape kits, is an objective method that has been established to be foundationally valid.

Validity as applied. Because errors due to human failures will dominate the chance of coincidental matches, the scientific criteria for validity as applied require that an expert (1) should have undergone rigorous and relevant proficiency testing to demonstrate their ability to reliably apply the method, (2) should routinely disclose in reports and testimony whether, when performing the examination, he or she was aware of any facts of the case that might influence the conclusion, and (3) should disclose, upon request, all information about quality testing and quality issues in his or her laboratory.

5.2 DNA Analysis of Complex-mixture Samples

Some investigations involve DNA analysis of complex mixtures of biological samples from multiple unknown individuals in unknown proportions. Such samples might arise, for example, from mixed blood stains. As DNA testing kits have become more sensitive, there has been growing interest in "touch DNA"—for example, tiny quantities of DNA left by multiple individuals on a steering wheel of a car.

Methodology

The fundamental difference between DNA analysis of complex-mixture samples and DNA analysis of single-source and simple mixtures lies not in the laboratory processing, but in the interpretation of the resulting DNA profile.

DNA analysis of complex mixtures—defined as mixtures with more than two contributors—is inherently difficult and even more for small amounts of DNA.[201] Such samples result in a DNA profile that superimposes multiple individual DNA profiles. Interpreting a mixed profile is different for multiple reasons: each individual may contribute two, one or zero alleles at each locus; the alleles may overlap with one another; the peak heights may differ considerably, owing to differences in the amount and state of preservation of the DNA from each source; and the "stutter peaks" that surround alleles (common artifacts of the DNA amplification process) can

[201] See, for example, SWGDAM document on interpretation of DNA mixtures. www.swgdam.org/#!public-comments/c1t82.

obscure alleles that are present or suggest alleles that are not present.[202] It is often impossible to tell with certainty which alleles are present in the mixture or how many separate individuals contributed to the mixture, let alone accurately to infer the DNA profile of each individual.[203]

Instead, examiners must ask: "Could a suspect's DNA profile be present *within* the mixture profile? And, what is the probability that such an observation might occur by chance?" The questions are challenging for the reasons given above. Because many different DNA profiles may fit within some mixture profiles, the probability that a suspect "cannot be excluded" as a possible contributor to complex mixture may be *much higher* (in some cases, millions of times higher) than the probabilities encountered for matches to single-source DNA profiles. As a result, proper calculation of the statistical weight is critical for presenting accurate information in court.

Subjective Interpretation of Complex Mixtures

Initial approaches to the interpretation of complex mixtures relied on subjective judgment by examiners, together with the use of simplified statistical methods such as the "Combined Probability of Inclusion" (CPI). These approaches are problematic because subjective choices made by examiners, such as about which alleles to include in the calculation, can dramatically alter the result and lead to inaccurate answers.

The problem with subjective analysis of complex-mixture samples is illustrated by a 2003 double-homicide case, *Winston v. Commonwealth*.[204] A prosecution expert reported that the defendant could not be excluded as a possible contributor to DNA on a discarded glove that contained a mixed DNA profile of at least three contributors; the defendant was convicted and sentenced to death. The prosecutor told the jury that the chance the match occurred by chance was 1 in 1.1 billion. A 2009 paper, however, makes a reasonable scientific case that that the chance is closer to 1 in 2—that is, 50 percent of the relevant population could not be excluded.[205] Such a large discrepancy is unacceptable, especially in cases where a defendant was sentenced to death.

Two papers clearly demonstrate that these commonly used approaches for DNA analysis of complex mixtures can be problematic. In a 2011 study, Dror and Hampikian tested whether irrelevant contextual information biased their conclusions of examiners, using DNA evidence from an actual adjudicated criminal case (a gang rape case in Georgia).[206] In this case, one of the suspects implicated another in connection with a plea bargain. The two experts who examined evidence from the crime scene were aware of this testimony against the suspect and knew that the plea bargain testimony could be used in court only with corroborating DNA evidence. Due to the

[202] Challenges with "low-template" DNA are described in a recent paper, Butler, J.M. "The future of forensic DNA analysis." *Philosophical Transactions of the Royal Society B,* 370: 20140252 (2015).

[203] See: Buckleton, J.S., Curran, J.M., and P. Gill. "Towards understanding the effect of uncertainty in the number of contributors to DNA stains." *Forensic Science International Genetics*, Vol. 1, No. 1 (2007): 20-8 and Coble, M.D., Bright, J.A., Buckleton, J.S., and J.M. Curran. "Uncertainty in the number of contributors in the proposed new CODIS set." *Forensic Science International Genetics*, Vol. 19 (2015): 207-11.

[204] *Winston v. Commonwealth,* 604 S.E.2d 21 (Va. 2004).

[205] Thompson, W.C. "Painting the target around the matching profile: the Texas sharpshooter fallacy in forensic DNA interpretation." *Law, Probability and Risk*, Vol. 8, No. 3 (2009): 257-76.

[206] Dror, I.E., and G. Hampikian. "Subjectivity and bias in forensic DNA mixture interpretation." *Science & Justice*, Vol. 51, No. 4 (2011): 204-8.

complex nature of the DNA mixture collected from the crime scene, the analysis of this evidence required judgment and interpretation on the part of the examiners. The two experts both concluded that the suspect could not be excluded as a contributor.

Dror and Hampikian presented the original DNA evidence from this crime to 17 expert DNA examiners, but without any of the irrelevant contextual information. They found that only 1 out of the 17 experts agreed with the original experts who were exposed to the biasing information (in fact, 12 of the examiners *excluded* the suspect as a possible contributor).

In another paper, de Keijser and colleagues presented 19 DNA experts with a mock case involving an alleged violent robbery outside a bar:

> *There is a male suspect, who denies any wrongdoing. The items that were sampled for DNA analysis are the shirt of the (alleged) female victim (who claims to have been grabbed by her assailant), a cigarette butt that was picked up by the police and that was allegedly smoked by the victim and/or the suspect, and nail clippings from the victim, who claims to have scratched the perpetrator.* [207]

Although all the experts were provided the same DNA profiles (prepared from the three samples above and the two people), their conclusions varied wildly. One examiner excluded the suspect as a possible contributor, while another examiner declared a match between the suspect's profile and a few minor peaks in the mixed profile from the nails—reporting a random match probability of roughly 1 in 209 million. Still other examiners declared the evidence inconclusive.

In the summer of 2015, a remarkable chain of events in Texas revealed that the problems with subjective analysis of complex DNA mixtures were not limited to a few individual cases: they were systemic.[208] The Texas Department of Public Safety (TX-DPS) issued a public letter on June 30, 2015 to the Texas criminal justice community noting that (1) the FBI had recently reported that it had identified and corrected minor errors in its population databases used to calculate statistics in DNA cases, (2) the errors were not expected to have any significant effect on results, and (2) the TX-DPS Crime Laboratory System would, upon request, recalculate statistics previously reported in individual cases.

When several prosecutors submitted requests for recalculation to TX-DPS and other laboratories, they were stunned to find that the statistics had changed dramatically—e.g., *from 1 in 1.4 billion to 1 in 36 in one case, from 1 in 4000 to inconclusive in another*. These prosecutors sought the assistance of the Texas Forensic Science Commission (TFSC) in understanding the reason for the change and the scope of potentially affected cases.

[207] de Keijser, J.W., Malsch, M., Luining, E.T., Kranenbarg, M.W., and D.J.H.M. Lenssen. "Differential reporting of mixed DNA profiles and its impact on jurists' evaluation of evidence: An international analysis." *Forensic Science International: Genetics*, Vol. 23 (2016): 71-82.

[208] Relevant documents and further details can be found at www.fsc.texas.gov/texas-dna-mixture-interpretation-case-review. Lynn Garcia, General Counsel for the Texas Forensic Science Commission, also provided a helpful summary to PCAST.

In consultation with forensic DNA experts, the TFSC determined that the large shifts observed in some cases were unrelated to the minor corrections in the FBI's population database, but rather were due to the fact that forensic laboratories had changed the way in which they calculated the CPI statistic—especially how they dealt with phenomena such as "allelic dropout" at particular DNA loci.

The TFSC launched a statewide DNA Mixture Notification Subcommittee, which included representatives of conviction integrity units, district and county attorneys, defense attorneys, innocence projects, the state attorney general, and the Texas governor. By September 2015, the TX-DPS had generated a county-by-county list of more than 24,000 DNA mixture cases analyzed from 1999-2015. Because TX-DPS is responsible for roughly half of the casework in the state, the total number of Texas DNA cases requiring review may exceed 50,000. (Although comparable efforts have not been undertaken in other states, the problem is likely to be national in scope, rather than specific to forensic laboratories in Texas.)

The TFSC also convened an international panel of scientific experts—from the Harvard Medical School, the University of North Texas Health Science Center, New Zealand's forensic research unit, and NIST—to clarify the proper use of CPI. These scientists presented observations at a public meeting, where many attorneys learned for the first time the extent to which DNA-mixture analysis involved subjective interpretation. Many of the problems with the CPI statistic arose because existing guidelines did not clearly, adequately, or correctly specify the proper use or limitations of the approach.

In summary, the interpretation of complex DNA mixtures with the CPI statistic has been an inadequately specified—and thus inappropriately subjective—method. As such, the method is clearly not foundationally valid.

In an attempt to fill this gap, the experts convened by TFSC wrote a joint scientific paper, which was published online on August 31, 2016.[209] The paper underscores the "pressing need . . . for standardization of an approach, training and ongoing testing of DNA analysts." The authors propose a set of specific rules for the use of the CPI statistic.

The proposed rules are clearly *necessary* for a scientifically valid method for the application of CPI. Because the paper appeared just as this report was being finalized, PCAST has not had adequate time to assess whether the rules are also *sufficient* to define an objective and scientifically valid method for the application of CPI.

Current Efforts to Develop Objective Methods

Given these problems, several groups have launched efforts to develop "probabilistic genotyping" computer programs that apply various algorithms to interpret complex mixtures. As of March 2014, at least 8 probabilistic genotyping software programs had been developed (called LRmix, Lab Retriever, likeLTD, FST, Armed Xpert, TrueAllele, STRmix, and DNA View Mixture Solution), with some being open source software and some being

[209] Bieber, F.R., Buckleton, J.S., Budowle, B., Butler, J.M., and M.D. Coble. "Evaluation of forensic DNA mixture evidence: protocol for evaluation, interpretation, and statistical calculations using the combined probability of inclusion." *BMC Genetics*. bmcgenet.biomedcentral.com/articles/10.1186/s12863-016-0429-7.

commercial products.[210] The FBI Laboratory began using the STRmix program less than a year ago, in December 2015, and is still in the process of publishing its own internal developmental validation.

These probabilistic genotyping software programs clearly represent a major improvement over purely subjective interpretation. However, they still require careful scrutiny to determine (1) whether the methods are scientifically valid, including defining the limitations on their reliability (that is, the circumstances in which they may yield unreliable results) and (2) whether the software correctly implements the methods. This is particularly important because the programs employ different mathematical algorithms and can yield different results for the same mixture profile.[211]

Appropriate evaluation of the proposed methods should consist of studies by multiple groups, *not associated with the software developers*, that investigate the performance and define the limitations of programs by testing them on a wide range of mixtures with different properties. In particular, it is important to address the following issues:

(1) How well does the method perform as a function of the number of contributors to the mixture? How well does it perform when the number of contributors to the mixture is *unknown*?

(2) How does the method perform as a function of the number of alleles shared among individuals in the mixture? Relatedly, how does it perform when the mixtures include related individuals?

(3) How well does the method perform—and how does accuracy degrade—as a function of the absolute and relative amounts of DNA from the various contributors? For example, it can be difficult to determine whether a small peak in the mixture profile represents a true allele from a minor contributor or a stutter peak from a nearby allele from a different contributor. (Notably, this issue underlies a current case that has received considerable attention.[212])

[210] The topic is reviewed in Butler, J.M. "Chapter 13: Coping with Potential Missing Alleles." *Advanced Topics in Forensic DNA Typing: Interpretation*. Waltham, MA: Elsevier/Academic, (2015): 333-48.

[211] Some programs use discrete (semi-continuous) methods, which use only allele information in conjunction with probabilities of allelic dropout and dropin, while other programs use continuous methods, which also incorporate information about peak height and other information. Within these two classes, the programs differ with respect to how they use the information. Some of the methods involve making assumptions about the number of individuals contributing to the DNA profile, and use this information to clean up noise (such as "stutter" in DNA profiles).

[212] In this case, examiners used two different DNA software programs (STRMix and TrueAllele) and obtained different conclusions concerning whether DNA from the defendant could be said to be included within the low-level DNA mixture profile obtained from a sample collected from one of the victim's fingernails. The judge ruled that the DNA evidence implicating the defendant was inadmissible. McKinley, J. "Potsdam Boy's Murder Case May Hinge on Minuscule DNA Sample From Fingernail." *New York Times.* See: www.nytimes.com/2016/07/25/nyregion/potsdam-boys-murder-case-may-hinge-on-statistical-analysis.html (accessed August 22, 2016). Sommerstein, D. "DNA results will not be allowed in Hillary murder trail." North Country Public Radio (accessed September 1, 2016). The decision can be found here: www.northcountrypublicradio.org/assets/files/08-26-16DecisionandOrder-DNAAnalysisAdmissibility.pdf.

(4) Under what circumstances—and why—does the method produce results (random inclusion probabilities) that differ substantially from those produced by other methods?

A number of papers have been published that analyze known mixtures in order to address some of these issues.[213] Two points should be noted about these studies. First, most of the studies evaluating software packages have been undertaken by the software developers themselves. While it is completely appropriate for method developers to evaluate their own methods, establishing scientific validity also requires scientific evaluation by other scientific groups that did not develop the method. Second, there have been few comparative studies across the methods to evaluate the differences among them—and, to our knowledge, no comparative studies conducted by independent groups.[214]

Most importantly, current studies have adequately explored only a limited range of mixture types (with respect to number of contributors, ratio of minor contributors, and total amount of DNA). The two most widely used methods (STRMix and TrueAllele) appear to be reliable within a certain range, based on the available evidence and the inherent difficulty of the problem.[215] Specifically, these methods appear to be reliable for three-person mixtures in which the minor contributor constitutes at least 20 percent of the intact DNA in the mixture and in which the DNA amount exceeds the minimum level required for the method.[216]

[213] For example: Perlin, M.W., Hornyak, J.M., Sugimoto, G., and K.W.P. Miller. "TrueAllele genotype identification on DNA mixtures containing up to five unknown contributors." *Journal of Forensic Sciences*, Vol. 60, No. 4 (2015): 857-868; Greenspoon S.A., Schiermeier-Wood L., and B.C. Jenkins. "Establishing the limits of TrueAllele® Casework: A validation study." *Journal of Forensic Sciences*. Vol. 60, No. 5 (2015):1263–76; Bright, J.A., Taylor, D., McGovern, C., Cooper, S., Russell, L., Abarno, D., and J.S. Buckleton. "Developmental validation of STRmix™, expert software for the interpretation of forensic DNA profiles." *Forensic Science International: Genetics*. Vol. 23 (2016): 226-39; Bright, J-A., Taylor D., Curran, J.S., and J.S. Buckleton. "Searching mixed DNA profiles directly against profile databases." *Forensic Science International: Genetics*. Vol. 9 (2014):102-10; Taylor D., Buckleton J, and I. Evett. "Testing likelihood ratios produced from complex DNA profiles." *Forensic Science International: Genetics.* Vol. 16 (2015): 165-171; Taylor D. and J.S. Buckleton. "Do low template DNA profiles have useful quantitative data?" *Forensic Science International: Genetics,* Vol. 16 (2015): 13-16.

[214] Bille, T.W., Weitz, S.M., Coble, M.D., Buckleton, J., and J.A. Bright. "Comparison of the performance of different models for the interpretation of low level mixed DNA profiles." *Electrophoresis*. Vol. 35 (2014): 3125–33.

[215] The interpretation of DNA mixtures becomes increasingly challenging as the number of contributors increases. See, for example: Taylor D., Buckleton J, and I. Evett. "Testing likelihood ratios produced from complex DNA profiles." *Forensic Science International: Genetics.* Vol. 16 (2015): 165-171; Bright, J.A., Taylor, D., McGovern, C., Cooper, S., Russell, L., Abarno, D., and J.S. Buckleton. "Developmental validation of STRmix™, expert software for the interpretation of forensic DNA profiles." *Forensic Science International: Genetics*. Vol. 23 (2016): 226-39; Bright, J-A., Taylor D., Curran, J.S., and J.S. Buckleton. "Searching mixed DNA profiles directly against profile databases." *Forensic Science International: Genetics*. Vol. 9 (2014):102-10; Bieber, F.R., Buckleton, J.S., Budowle, B., Butler, J.M., and M.D. Coble. "Evaluation of forensic DNA mixture evidence: protocol for evaluation, interpretation, and statistical calculations using the combined probability of inclusion." *BMC Genetics*. bmcgenet.biomedcentral.com/articles/10.1186/s12863-016-0429-7.

[216] Such three-person samples involving similar proportions are more straightforward to interpret owing to the limited number of alleles and relatively similar peak height. The methods can also be reliably applied to single-source and simple-mixture samples, provided that, in cases where the two contributions cannot be separated by differential extraction, the proportion of the minor contributor is not too low (e.g., at least 10 percent).

For more complex mixtures (e.g. more contributors or lower proportions), there is relatively little published evidence.[217] In human molecular genetics, an experimental validation of an important diagnostic method would typically involve hundreds of distinct samples.[218] One forensic scientist told PCAST that many more distinct samples have, in fact, been analyzed, but that the data have not yet been collated and published.[219] Because empirical evidence is essential for establishing the foundational validity of a method, PCAST urges forensic scientists to submit and leading scientific journals to publish high-quality validation studies that properly establish the range of reliability of methods for the analysis of complex DNA mixtures.

When further studies are published, it will likely be possible to extend the range in which scientific validity has been established to include more challenging samples. As noted above, such studies should be performed by or should include independent research groups not connected with the developers of the methods and with no stake in the outcome.

Conclusion

Based on its evaluation of the published literature to date, PCAST reached several conclusions concerning the foundational validity of methods for the analysis of complex DNA mixtures. We note that foundational validity must be established with respect to a specified method applied to a specified range. In addition to forming its own judgment, PCAST also consulted with John Butler, Special Assistant to the Director for Forensic Science at NIST and Vice Chair of the NCFS.[220] Butler concurred with PCAST's finding.

[217] For four-person mixtures, for example, papers describing experimental validations with known mixtures using TrueAllele involve 7 and 17 distinct mixtures, respectively, with relatively large amounts of DNA (at least 200 pg), while those using STRMix involve 2 and 3 distinct mixtures, respectively, but use much lower amounts of DNA (in the range of 10 pg). Greenspoon S.A., Schiermeier-Wood L., and B.C. Jenkins. "Establishing the limits of TrueAllele® Casework: A validation study." *Journal of Forensic Sciences*. Vol. 60, No. 5 (2015):1263–76; Perlin, M.W., Hornyak, J.M., Sugimoto, G., and K.W.P. Miller. "TrueAllele genotype identification on DNA mixtures containing up to five unknown contributors." *Journal of Forensic Sciences*, Vol. 60, No. 4 (2015): 857-868; Taylor, D. "Using continuous DNA interpretation methods to revisit likelihood ratio behavior." *Forensic Science International: Genetics,* Vol. 11 (2014): 144-153; Taylor D., Buckleton J, and I. Evett. "Testing likelihood ratios produced from complex DNA profiles." *Forensic Science International: Genetics.* Vol. 16 (2015): 165-171; Taylor D. and J.S. Buckleton. "Do low template DNA profiles have useful quantitative data?" *Forensic Science International: Genetics,* Vol. 16 (2015): 13-16; Bright, J.A., Taylor, D., McGovern, C., Cooper, S., Russell, L., Abarno, D., J.S. Buckleton. "Developmental validation of STRmix^{TM}, expert software for the interpretation of forensic DNA profiles." *Forensic Science International: Genetics*. Vol. 23 (2016): 226-39.

[218] Preparing and performing PCR amplication on hundreds of DNA mixtures is straightforward; it can be accomplished within a few weeks or less.

[219] PCAST interview with John Buckleton, Principal Scientist at New Zealand's Institute of Environmental Science and Research and a co-developer of STRMix.

[220] Butler is a world authority on forensic DNA analysis, whose Ph.D. research, conducted at the FBI Laboratory, pioneered techniques of modern forensic DNA analysis and who has written five widely acclaimed textbooks on forensic DNA typing. See: Butler, J.M. *Forensic DNA Typing: Biology and Technology behind STR Markers*. Academic Press, London (2001); Butler, J.M. *Forensic DNA Typing: Biology, Technology, and Genetics of STR Markers (2nd Edition)*. Elsevier Academic Press, New York (2005); Butler, J.M. *Fundamentals of Forensic DNA Typing.* Elsevier Academic Press, San Diego (2010); Butler, J.M. *Advanced Topics in Forensic DNA Typing: Methodology.* Elsevier Academic Press, San Diego (2012); Butler, J.M. *Advanced Topics in Forensic DNA Typing: Interpretation*. Elsevier Academic Press, San Diego (2015).

Finding 3: DNA analysis of complex-mixture samples

Foundational validity. PCAST finds that:

(1) Combined-Probability-of-Inclusion (CPI)-based methods. DNA analysis of complex mixtures based on CPI-based approaches has been an inadequately specified, subjective method that has the potential to lead to erroneous results. As such, it is not foundationally valid.

A very recent paper has proposed specific rules that address a number of problems in the use of CPI. These rules are clearly *necessary*. However, PCAST has not adequate time to assess whether they are also *sufficient* to define an objective and scientifically valid method. If, for a limited time, courts choose to admit results based on the application of CPI, validity as applied would require that, at a minimum, they be consistent with the rules specified in the paper.

DNA analysis of complex mixtures should move rapidly to more appropriate methods based on probabilistic genotyping.

(2) Probabilistic genotyping. Objective analysis of complex DNA mixtures with probabilistic genotyping software is relatively new and promising approach. Empirical evidence is required to establish the foundational validity of each such method within specified ranges. At present, published evidence supports the foundational validity of analysis, with some programs, of DNA mixtures of 3 individuals in which the minor contributor constitutes at least 20 percent of the intact DNA in the mixture and in which the DNA amount exceeds the minimum required level for the method. The range in which foundational validity has been established is likely to grow as adequate evidence for more complex mixtures is obtained and published.

Validity as applied. For methods that are foundationally valid, validity as applied involves similar considerations as for DNA analysis of single-source and simple-mixtures samples, with a special emphasis on ensuring that the method was applied correctly and within its empirically established range.

The Path Forward

There is a clear path for extending the range over which objective methods have been established to be foundationally valid—specifically, through the publication of appropriate scientific studies.

Such efforts will be aided by the creation and dissemination (under appropriate data-use and data-privacy restrictions) of large collections of hundreds of DNA profiles created from known mixtures—representing widely varying complexity with respect to (1) the number of contributors, (2) the relationships among contributors, (3) the absolute and relative amounts of materials, and (4) the state of preservation of materials—that can be used by independent groups to evaluate and compare the methods. Notably, the PROVEDIt Initiative (Project Research Openness for Validation with Experimental Data) at Boston University has made available a resource of

25,000 profiles from DNA mixtures.[221,222] In addition to scientific studies on common sets of samples for the purpose of evaluating foundational validity, individual forensic laboratories will want to conduct their own internal developmental validation studies to assess the validity of the method in their own hands.[223]

NIST should play a leadership role in this process, by ensuring the creation and dissemination of materials and stimulating studies by independent groups through grants, contracts, and prizes; and by evaluating the results of these studies.

5.3 Bitemark Analysis

Methodology

Bitemark analysis is a subjective method. It typically involves examining marks left on a victim or an object at the crime scene, and comparing those marks with dental impressions taken from a suspect.[224] Bitemark comparison is based on the premises that (1) dental characteristics, particularly the arrangement of the front teeth, differ substantially among people and (2) skin (or some other marked surface at a crime scene) can reliably capture these distinctive features.

Bitemark analysis begins with an examiner deciding whether an injury is a mark caused by human teeth.[225] If so, the examiner creates photographs or impressions of the questioned bitemark and of the suspect's dentition; compares the bitemark and the dentition; and determines if the dentition (1) cannot be excluded as having made the bitemark, (2) can be excluded as having made the bitemark, or (3) is inconclusive. The bitemark standards do not provide well-defined standards concerning the degree of similarity that must be identified to support a reliable conclusion that the mark could have or could not have been created by the dentition in question. Conclusions about all these matters are left to the examiner's judgment.

Background Studies

Before turning to the question of foundational validity, we discuss some background studies (concerning such topics as uniqueness and consistency) that shed some light on the field. These studies cast serious doubt on the fundamental premises of the field.

[221] See: www.bu.edu/dnamixtures.
[222] The collection contains DNA samples with 1- to 5-person DNA mixtures, amplified with targets ranging from 1 to 0.007 ng. In the multi-person mixtures, the ratio of contributors range from 1:1 to 1:19. Additionally, the profiles were generated using a variety of laboratory conditions from samples containing pristine DNA; UV damaged DNA; enzymatically or sonically degraded DNA; and inhibited DNA.
[223] The FBI Laboratory has recently completed a developmental validation study and is preparing it for publication.
[224] Less frequently, marks are found on a suspected perpetrator that may have come from a victim.
[225] ABFO Bitemark Methodology Standards and Guidelines, abfo.org/wp-content/uploads/2016/03/ABFO-Bitemark-Standards-03162016.pdf (accessed July 2, 2016).

Exhibit 2

SWGDAM Guidelines for Validation of Probabilistic Genotyping Systems

Scientific Working Group on DNA Analysis Methods

Guidelines for the Validation of Probabilistic Genotyping Systems

Table of Contents

SWGDAM Guidelinesfor the Validation of Probabilistic Genotyping Systems

The Scientific Working Group on DNA Analysis Methods, better known by its acronym of SWGDAM, is a group of approximately 50 scientists representing Federal, State, and Local forensic DNA laboratories in the United States and Canada. During meetings, which are held twice a year, Committees discuss topics of interest to the forensic DNA community and often develop documents to provide direction and guidance for the community. In some instances, an Ad Hoc Working Group may be empanelled to address a particular topic outside of the routine SWGDAM January/July meeting schedule.These Guidelines,drafted by the SWGDAM Ad Hoc Working Group on Probabilistic Genotyping,wereapproved by the SWGDAM Executive Board for public comment in March 2015.Following the public comment period, the Ad Hoc Working Group forwarded the Final Guidelines to the SWGDAM Executive Board and they were approved for posting on the SWGDAM web site on June 15, 2015.

Guidance is provided herein for the validation of probabilistic genotyping software used for the analysis of autosomal short tandem repeat (STR) typing results. These guidelines are not intended to be applied retroactively. It is anticipated that they will evolve with future developments in probabilistic genotyping systems.

Introduction

Probabilistic genotyping refers to the use of biological modeling, statistical theory, computer algorithms, and probability distributions to calculate likelihood ratios (LRs) and/or infer genotypes for the DNA typing results of forensic samples (“forensic DNA typing results”). Human interpretation and review is required for the interpretation of forensic DNA typing results in accordance with the FBI Director’s Quality Assurance Standards for Forensic DNA Testing Laboratories[1]. Probabilistic genotyping is a tool to assist the DNA analyst in the interpretation of forensic DNA typing results. Probabilistic genotyping is not intended to replace the human evaluation of the forensic DNA typing results or the human review of the output prior to reporting.

A probabilistic genotyping system is comprised of software, or software and hardware, with analytical and statistical functions that entail complex formulae and algorithms. Particularly useful for low-level DNA samples (i.e., those in which the quantity of DNA for individuals is such that stochastic effects may be observed) and complex mixtures(i.e., multi-contributor samples, particularly those exhibiting allele sharing and/or stochastic effects), probabilistic genotyping approaches can reduce subjectivity in the analysis of DNA typing results.Historical methods of mixture interpretation consider all interpreted genotype combinations to be equally probable, whereas probabilistic approaches provide a statistical weighting to the different genotype combinations. Probabilistic genotyping does not utilize a stochastic threshold. Instead, it incorporates a probability of alleles dropping out or in.In making use of more genotyping information when performing statistical calculations and evaluating potential DNA contributors, probabilistic genotyping enhances the ability to distinguish true contributors and non-contributors.A higher LR is typically obtained when evaluating a person of interest (POI) who is a true contributor to the evidence profile, and a lower LR is typically obtained when the POI is not a true contributor. While the absence of an allele or the presence of additional allele(s)

[1] Probabilistic genotyping is to be distinguished from an Expert System.An Expert System, if NDIS approved and properly validated in accordance with the QAS, may only be used by a laboratory on database, known or casework reference samples to replace the manual review in accordance with the QAS and NDIS Operational Procedures. Expert Systems are not approved for use on forensic or forensic mixture DNA samples.

relative to a reference sample may support an exclusion, probabilistic genotyping approaches allow inclusion and exclusion hypotheses to be consideredby calculating a LR in which allele drop-out and drop-in may be incorporated.

The use of a likelihood ratio as a reporting statistic for probabilistic genotypingdiffers substantially from binary statistics such as the combined probability of exclusion. Prior to validating a probabilistic genotyping system, the laboratory should ensure that it possesses the appropriate foundational knowledge in the calculation and interpretation of likelihood ratios. Laboratories should also be aware of the features and limitations of various probabilistic genotyping programs and the impact that those items will have on the validation process. Depending on the performance characteristics of the software, prerequisite studies may be required to, for example, establish parameters forallele drop-out and drop-in, stutter expectations, peak height variation, and the number of contributors to a mixture.Each laboratory seeking to evaluate a probabilistic genotyping system must determine which validation studies are relevant to the methodology, in the context of its application, to demonstrate the reliability of the system and any potential limitations. The laboratory must determine the number of samples required to satisfy each guideline and may determine that a study is not necessary. Some studies described herein may also be suitable for evaluating material modifications to existing procedures.

Background

Please refer to the SWGDAM Validation Guidelines for DNA Analysis Methods andthe FBI Quality Assurance Standards for Forensic DNA Testing Laboratories and for DNA Databasing Laboratories (QAS) for general background information regarding validation and definition of terms.

Probabilistic genotyping may generate a number of possible genotype combinations for a given profile, where some genotypes may be assigned more weight than others. Allele drop-in and drop-out probabilities may be used in the determination of the weights associated with each of the possible genotypes. There are two main approaches to probabilistic genotyping: the semi-continuous method and fully continuous method. The semi-continuous method focuses only on the alleles present in the profile and addresses all possible genotype combinations of the observed alleles in conjunction with a probability of drop-out and drop-in. Analysis parameters such as peak height variation, mixture ratios and stutter percentages are not typically utilized by semi-continuous software systems, although these elements may be considered during the initial manual evaluation of the data. The fully continuous method generally utilizesmore of the biological information in the profile, such as peak heights, stutter percentages and mixture ratios. The weighting of genotype combinations as more or less probable may be inferred from the data through methods such as Markov Chain Monte Carlo (MCMC) samplings from probability distributions.

The analyst will need to employ some level of interpretation before using the software to perform the calculations and should visually interpret allelic and non-allelic peaks and other characteristics of the DNA typing results, as necessitated by the software. For example, the analyst may be required to estimate and use a specific number of contributors in a statistical calculation when interpreting a DNA mixture, or to assess whether typing results should be interpreted or not based on quality.

Forensic DNA typing results interpreted by a DNA analyst using probabilistic genotyping software may be eligible for CODIS entry and upload to NDIS in accordance with the NDIS Operational Procedures if the probabilistic genotyping software has been properly validated pursuant to the QAS and these Guidelines.

1. Validation of Probabilistic Genotyping Systems

1.1. The laboratory shall validate a probabilistic genotyping system prior to usage for forensic applications.

1.2. The laboratory shall document all validation studies in accordance with the FBI Quality Assurance Standards for Forensic DNA Testing Laboratories.

1.3. The laboratory should document or have access to documentation that explains how the software performs its operations and activities, to include the methods of analysis and statistical formulae, the data to be entered in the system, the operations performed by each portion of the user interface, the workflow of the system, and the system reports or other outputs. This information enables the laboratory to identify aspects of the system that should be evaluated through validation studies.

2. **System control**

2.1. The laboratory should verify that the software is installed on computers suited to run the software, that the system has been properly installed, and that the configurations are correct.

2.2. The laboratory should, where possible, ensure the following system control measures are in effect:

2.2.1. Every software release should have a unique version number. This version number should be referenced in any validation documentation or published results.

2.2.2. Appropriate security protection to ensure only authorized users can access the software and data.

2.2.3. Audit trails to track changes to system data and/or verification of system settings in place each time a calculation is run.

2.2.4. User-level security to ensure that system users only perform authorized actions.

3. **Developmental Validation**

Developmental validation of a probabilistic genotyping system is the acquisition of test data to verify the functionality of the system, the accuracy of statistical calculations and other results, the appropriateness of analytical and statistical parameters, and the determination of limitations. Developmental validation may be conducted by the manufacturer/developer of the application or the testing laboratory. Developmental validation should also demonstrate any known or potential limitations of the system.

3.1. The underlying scientific principle(s) of the probabilistic genotyping methods and characteristics of the software should be published in a peer-reviewed scientific journal. The underlying scientific principles of probabilistic genotyping include, but are not limited to, modeling of stutter, allelic drop-in and drop-out, Bayesian prior assumptions such as allele probabilities, and statistical formulae used in the calculation and algorithms.

3.2. Developmental validation should address, where applicable, the following:

3.2.1. Sensitivity – Studies should assess the ability of the system to reliably determine the presence of a contributor's(s') DNA over a broad variety of evidentiary typing results (to include mixtures and low-level DNA quantities). This should be evaluated using various sample types (e.g., different numbers of contributors, mixture proportions, and template quantities).

3.2.1.1. Sensitivity studies should demonstrate the potential for Type I errors (i.e., incorrect rejection of a true hypothesis), in which, for example, a contributor fails to yield a LR greater than 1 and thus his/her presence in the mixture is not supported.

3.2.1.2. Sensitivity studies should demonstrate the range of LR values that can be expected for contributors.

3.2.2. Specificity – Studies should evaluate the ability of the system to provide reliable results for non-contributors over a broad variety of evidentiary typing results (to include mixtures and low-level DNA quantities). This should be evaluated using various sample types (e.g., different numbers of contributors, mixture proportions, and template quantities).

3.2.2.1. Specificity studies should demonstrate the potential for Type II errors (i.e., failure to reject a false hypothesis), in which, for example, a non-contributor yields a LR greater than 1 and thus his/her presence in the mixture is supported.

3.2.2.2.Specificity studies should demonstrate the range of LR values that can be expected for non-contributors.

3.2.3. Precision – Studies should evaluate the variation in Likelihood Ratios calculated from repeated software analyses of the same input data. This should be evaluated using various sample types (e.g., different numbers of contributors, mixture proportions, and template quantities).

3.2.3.1. Some probabilistic genotyping approaches may not produce the same LR from repeat analyses. Where applicable, these studies should therefore demonstrate the range of LR values that can be expected from multiple analyses of the same data and are the basis for establishing an acceptable amount of variation in LRs.

3.2.3.2. Any parameter settings (e.g., iterations of the MCMC)that can reduce variability should be evaluated. For example, for some complex mixtures (e.g., partial profiles with more than three contributors), increasing the number of MCMC iterations can reduce variation in the likelihood ratio.

3.2.4. Case-type Samples – Studies should assess a range of data types exhibiting features that are representative of those typically encountered by testing laboratories. These features include those derived from mixtures and single-source samples, such as stutter, masked/shared alleles, differential and preferential amplification, degradation and inhibition.

3.2.4.1. These studies should demonstrate sample and/or data types that can be reliably evaluated using the probabilistic genotyping system.

3.2.5. Control Samples – If the software is designed to assess controls, studies should evaluate whether correct results are obtained with control samples.

3.2.6. Accuracy – Studies should assess the accuracy of the calculations performed by the system, as well as allele designation functions, where applicable.

3.2.6.1. These studies should include the comparison of the results produced by the probabilistic genotyping software to manual calculations, or results produced with an alternate software program or application, to aid in assessing accuracy of results generated by the probabilistic genotyping system. Calculations of some profiles (e.g., complex mixtures), however, may not be replicable outside of the probabilistic genotyping system.

3.2.6.2. If the software uses raw data files from a genetic analyzer as input data, the peak calling, sizing and allele designation functions should be compared to the results of another software system to assess accuracy. Allele designations should also be compared to known genotypes where available.

4. Internal Validation

Internal validation of a probabilistic genotyping software system is the accumulation of test data within the laboratory to demonstrate that the established parameters, software settings, formulae, algorithms and functions perform as expected. In accordance with the QAS, internal validation data may be shared by all locations in a multi-laboratory system.

Depending on the features and capabilities of the probabilistic genotyping system, some DNA typing results may or may not be determined to be suitable for such analysis. To identify data features (e.g., minimum quality requirements, number of contributors) that render a profile appropriate or inappropriate for probabilistic genotyping, the laboratory should test data across a range of characteristics that are representative of those typically encountered by the testing laboratory. Data should be selected to test the system's capabilities and to identify its limitations. In particular, complex mixtures and low-level contributors should be evaluated thoroughly during internal validation, as the data from such samples generally help to define the software's limitations, as well as sample and/or data types which may potentially not be suitable for computer analysis. In addition, some exclusions may be evident without the aid of probabilistic software.

If conducted within the same laboratory, developmental validation studies may satisfy some of the elements of the internal validation guidelines.

4.1. The laboratory should test the system using representative data generated in-house with the amplification kit, detection instrumentation and analysis software used for casework. Additionally, some studies may be conducted by using artificially created or altered input files to further assess the capabilities and limitations of the software. Internal validation should address, where applicable to the software being evaluated:

4.1.1. Specimens with known contributors, as well as case-type specimens that may include unknown contributors.

4.1.2. Hypothesis testing with contributors and non-contributors

4.1.2.1. The laboratory should evaluate more than one set of hypotheses for individual evidentiary profiles to aid in the development of policies regarding the formulation of hypotheses. For example, if there are two persons of interest, they may be evaluated as co-contributors and, alternatively, as each contributing with an unknown individual. The hypotheses used for evaluation of casework profiles can have a significant impact on the results obtained.

4.1.3. Variable DNA typing conditions (e.g., any variations in the amplification and/or electrophoresis parameters used by the laboratory to increase or decrease the detection of alleles and/or artifacts)

4.1.4. Allelic peak height, to include off-scale peaks

4.1.5. Single-source specimens

4.1.6. Mixed specimens

4.1.6.1. Various contributor ratios (e.g., 1:1 through 1:20, 2:2:1, 4:2:1, 3:1:1, etc.)

4.1.6.2. Various total DNA template quantities

4.1.6.3. Various numbers of contributors. The number of contributors evaluated should be based on the laboratory'sintended use of the software.A range of contributor numbers should be evaluated in order to define the limitations of the software.

4.1.6.4. If the number of contributors is input by the analyst, both correct and incorrect values (i.e., over- and under-estimating) should be tested.

4.1.6.5. Sharing of alleles among contributors

4.1.7. Partial profiles, to include the following:

4.1.7.1. Allele and locus drop-out

4.1.7.2. DNA degradation

4.1.7.3. Inhibition

4.1.8. Allele drop-in

4.1.9. Forward and reverse stutter

4.1.10. Intra-locus peak height variation

4.1.11. Inter-locus peak height variation

4.1.12. For probabilistic genotyping systems that require in-house parameters to be established, the internal validation tests should be performed using those same parameters. The data set used to establish the parameters should be different from the data set used to validate the software using those parameters.

4.1.13. Sensitivity, specificity and precision, as described for Developmental Validation

4.1.14. Additional challenge testing (e.g., the inclusion of non-allelic peaks such as bleed-through and spikes in the typing results)

4.2. Laboratories with existing interpretation procedures should compare the results of probabilistic genotyping and of manual interpretation of the same data, notwithstanding the fact that probabilistic genotyping is inherently different from and not directly comparable to binary interpretation.The weights of evidence that are generated by these two approaches are based on different assumptions, thresholds and formulae. However, such a comparison should be conducted and evaluated for general consistency.

4.2.1. The laboratory should determine whether the results produced by the probabilistic genotyping software are intuitive and consistent with expectations based on non-probabilistic mixture analysis methods.

4.2.1.1. Generally, known specimens that are included based on non-probabilistic analyses would be expected to also be included based on probabilistic genotyping.

4.2.1.2. For single-source specimens with high quality results, genotypes derived from non-probabilistic analyses of profiles above the stochastic threshold should be in complete concordance with the results of probabilistic methods.

4.2.1.3. Generally, as the analyst’s ability to deconvolute a complex mixture decreases, so do the weightings of individual genotypes within a set determined by the software.

5. Modification to Software

Modification to probabilistic genotyping software shall be addressed in accordance with the QAS.

5.1. Modification to the system such as a hardware or software upgrade that does not impact interpretation or analysis of the typing results or the statistical analysis shall require a performance check prior to implementation.

5.2. A significant change(s) to the software, defined as that which may impact interpretation or the analytical process, shall require validation prior to implementation.

5.3. Data used during the initial validation may be re-evaluated as a performance check or for subsequent validation assessment. The laboratory must determine the number and type of samples required to establish acceptable performance in consideration of the software modification.

References and Suggested Readings

Federal Bureau of Investigation, (2015) *NDIS Operational Procedures Manual*, available at http://www.fbi.gov/about-us/lab/biometric-analysis/codis/ndis-procedures-manual.

Federal Bureau of Investigation (2011) *Quality Assurance Standards for Forensic DNA Testing Laboratories*, available at http://www.fbi.gov/about-us/lab/codis/qas-standards-for-forensic-dna-testing-laboratories-effective-9-1-2011

Gill, P. et al. (2012)*DNA Commission of the International Society of Forensic Genetics: Recommendations on the evaluation of STR typing results that may include drop-out and/or drop-in using probabilistic methods*, Forensic Science International Genetics 6(6): 679-688.

Kelly. H. et al. (2014)*A comparison of statistical models for the analysis of complex forensic DNA profiles*, Science &Justice 54(1): 66-70.

Scientific Working Group on DNA Analysis Methods (2010) *Interpretation Guidelines for Autosomal STR Typing by Forensic DNA Testing Laboratories*, available at http://www.swgdam.org/Interpretation_Guidelines_January_2010.pdf.

Scientific Working Group on DNA Analysis Methods (2012) *Validation Guidelines for DNA Analysis Methods*, available at http://www.swgdam.org/SWGDAM_Validation_Guidelines_APPROVED_Dec_2012.pdf .

Scientific Working Group on DNA Analysis Methods (2014) *Guidelines for STR Enhanced Detection Methods*, available at http://swgdam.org/SWGDAM%20Guidelines%20for%20STR%20Enhanced%20Detection%20Methods%20FINAL%20100614.pdf.

Steele, C. D. and Balding, D. J. (2014) *Statistical Evaluation of Forensic DNA Profile Evidence,* Annu. Rev. Stat. Appl. 1:361-384.

Exhibit 3

A summary of the seven identified miscodes in STRmix

A summary of the seven identified miscodes in STRmix™

Number	Description	Effect
1	Corrected a minor miscoding of the Balding and Nichols formulae, under the following two conditions, theta values are above 0, and there are unknown contributors under the H_p hypothesis	Present in versions up to, but not including, V2.0. It affected the *LR* in minor way, with revised *LR*s generally within one order of magnitude. It was detected by a third party laboratory repeating calculations by hand
2	The miscode affected the numerical value of the *LR* in rare instances where a three or four person mixture is interpreted with one assumed contributor, one POI and where one of the unknown contributors shares a genotype with the assumed.	This miscode was discovered after a case was brought to Forensic Science South Australia's attention in December 2014 by Queensland Health.[1] Present in STRmix™ versions up to but not including V2.0.6. All potentially affected cases in Australia and New Zealand were examined and very few instances of a different *LR* occurred outside Queensland. To get an effect from this miscode a specific set of circumstances was needed which are unlikely to occur and which had not been specifically tested in developmental validation prior to this date of discovery. Revised *LR*s were generally within one order of magnitude smaller
3	Involved rare cases with multiple dropped alleles across all contributors within one locus	This error was found to be due to a change in STRmix™ that was introduced in the V2.3 series and has been present in all versions subsequent to this up to V2.3.09 and V2.4.04 (inclusive). The effect on *LR*s was in a conservative direction and very minor.
4	This miscode affected only the database search *LR* (assuming the presence of forward stutter peaks and/or multiple drop-in alleles, and modelling of these artifacts were enabled) and to the Highest Posterior Density (HPD) calculations	Present in V1.0 to V2.4.06 for drop-in modelling and V2.4 to V2.4.06 for forward stutter modelling. There was no detectable effect on the *LR* in 92 profiles tested.
5	A change was made to the *LR* calculation (unrelated point estimate, stratified, unified and HPD) for mixed DNA profiles when there are multiple	Present in all versions preceding V2.4.08. Changes in the *LR* were usually less than one order of magnitude.

[1] http://www.couriermail.com.au/news/queensland/queensland-authorities-confirm-miscode-affects-dna-evidence-in-criminal-cases/news-story/833c580d3f1c59039efd1a2ef55af92b

The number of cases affected is 23 not 60 see
https://www.health.qld.gov.au/__data/assets/pdf_file/0029/633368/dohdl1617012.pdf

We do not have good data on the details of these cases. Queensland Health shared data up to and including the first of the set of profiles associated with this, then ceased citing initially that sharing of profiles was now illegal. Hence we do not know the magnitude of any change. Subsequently Queensland Health have informed us that they can share the profiles but this has never occurred. We speculate that the much higher rate of exposure of this miscode in Queensland is a result of their workflow which, we believe, in some instances adds an extra contributor beyond that required at a minimum to explain the profile. This exposed a part of the code with an error in it. This usage had not been envisaged and had not be tested prior to this event.

	contributors considered under H_p who are unknown under H_d. The contributors must have dropped alleles (either the same or different) at the same locus	
6	A change was been made to the way drop-in alleles are assigned within the determination of the genotype array within the pre burn-in phase. This miscode affected all *LR* calculations including within the database search (standard and familial *LR*s), unrelated and related scenarios	Present in all versions preceding V2.4.08. Changes in the *LR* were usually less than one order of magnitude
7	A minor anomaly in the familial search *LR* was identified. The issue affected mixed and single source profiles with dropout where on some occasions an incorrect allele frequency was assigned to the alleles	Present in all versions preceding V2.4.08. The comparison of *LR*s between versions indicated that this error was mostly within one order of magnitude

Exhibit 4

Frederick R. Bieber, et al., Evaluation of forensic DNA mixture evidence: protocol for evaluation, interpretation, and statistical calculations using the combined probability of inclusion

Bieber *et al. BMC Genetics* (2016) 17:125
DOI 10.1186/s12863-016-0429-7

BMC Genetics

METHODOLOGY ARTICLE

Open Access

Evaluation of forensic DNA mixture evidence: protocol for evaluation, interpretation, and statistical calculations using the combined probability of inclusion

Frederick R. Bieber[1*†], John S. Buckleton[2,3†], Bruce Budowle[4†], John M. Butler[5] and Michael D. Coble[6]

Abstract

Background: The evaluation and interpretation of forensic DNA mixture evidence faces greater interpretational challenges due to increasingly complex mixture evidence. Such challenges include: casework involving low quantity or degraded evidence leading to allele and locus dropout; allele sharing of contributors leading to allele stacking; and differentiation of PCR stutter artifacts from true alleles. There is variation in statistical approaches used to evaluate the strength of the evidence when inclusion of a specific known individual(s) is determined, and the approaches used must be supportable. There are concerns that methods utilized for interpretation of complex forensic DNA mixtures may not be implemented properly in some casework. Similar questions are being raised in a number of U.S. jurisdictions, leading to some confusion about mixture interpretation for current and previous casework.

Results: Key elements necessary for the interpretation and statistical evaluation of forensic DNA mixtures are described. Given the most common method for statistical evaluation of DNA mixtures in many parts of the world, including the USA, is the Combined Probability of Inclusion/Exclusion (CPI/CPE). Exposition and elucidation of this method and a protocol for use is the focus of this article. Formulae and other supporting materials are provided.

Conclusions: Guidance and details of a DNA mixture interpretation protocol is provided for application of the CPI/CPE method in the analysis of more complex forensic DNA mixtures. This description, in turn, should help reduce the variability of interpretation with application of this methodology and thereby improve the quality of DNA mixture interpretation throughout the forensic community.

Keywords: Forensic DNA mixtures, Combined probability of inclusion, CPI, Allele drop-out, Stochastic threshold

Abbreviations: AT, Analytical threshold; CE, Capillary electrophoresis; CPE, Combined probability of exclusion; CPI, Combined probability of inclusion; LR, Likelihood ratio; MAC, Minimum allele contribution; mtDNA, Mitochondrial DNA; PCR, Polymerase chain reaction; POI, Person of interest; RFU, Relative fluorescent unit; RMP, Random match probability; SF, Stutter filter value; SPH, Peak height value in the stutter position; STRmix^{TM}, Forensic software; STR, Short tandem repeat; ST, Stochastic threshold; SWGDAM, Scientific working group on DNA analysis methods

* Correspondence: frbieber@bics.bwh.harvard.edu
†Equal contributors
[1]Center for Advanced Molecular Diagnostics, Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, 75 Francis Street, Boston, MA 02115, USA
Full list of author information is available at the end of the article

Background

Forensic DNA Mixtures

A DNA mixture refers to a biological sample that originated from two or more donors and is determined after a DNA profile is generated. Mixture evidence has always been a part of casework; however there are indications that the fraction of samples presenting as a mixture may have increased, presumably due to changes in methodology, sampling strategies, types of cases (e.g., high volume crime). A retrospective study over a 4 year period in Spain [1] found, in the early years of short tandem repeat (STR) typing, that nearly 95 % of casework samples produced single-source profiles. Initially most mixtures were derived from sexual assault evidence, fingernail cuttings taken by police or at autopsy, from products of conception, and other similar materials. Mixtures from such evidence, combined with the sensitivity of detection of kits at that time, commonly involved only two contributors and one of them (e.g., in sexual assault evidence the person from whom the sample was obtained; in products of conception the biological mother) was "known" and the remaining part of the DNA mixture profile could be inferred to have originated from the second person (i.e., possible person of interest or foreign contributor) [2]. Evaluation of such evidence is, accordingly, comparatively straightforward as the amount of DNA is typically ample, contributions from different individuals are readily evaluated, and the allelic contributions to the DNA evidence of the known individual can be easily "subtracted" from the DNA mixture profile.

In current forensic investigations DNA mixtures occur commonly [3]. Moreover, crime laboratories are being asked to evaluate many more poor-quality, low-template, and complex DNA mixtures. In addition, the forensic community is utilizing methods with an increased sensitivity of detection due to improvements in DNA extraction methods, enhanced multiplex kits, and use of increased number of PCR cycles (or other manipulations) which in turn enable analysis of more challenging and complex mixtures.

At this time, the most commonly used method for forensic evaluation of DNA evidence is the assessment of short tandem repeat (STR) polymorphisms present at multiple distinct genetic loci [4–6]. The amplified products are separated by size using capillary electrophoresis (CE). DNA sequencing also can be used for assessment of STR alleles as well as mtDNA types [7, 8]. After STR analysis, the presence of three or more allelic peaks at two or more genetic loci or peak height differences that are greater than a defined heterozygote peak height ratio are indications that multiple donors contributed to the specific tested DNA sample. A "complex DNA mixture" may contain more than two donors, one or more of the donors may have contributed a low amount of DNA template, or the sample may be somewhat degraded. Low amounts of input DNA will present random (stochastic) effects during DNA amplification on results of STR testing which in turn can lead to failure to detect some or all of the alleles of a true donor (i.e., allele drop-out) [9, 10].

The combined probability of inclusion (CPI) [3, 11, 12] is the most commonly used method in the Americas, Asia, Africa, and the Middle East to assign the weight of evidence where a probative profile is obtained from an evidentiary sample. The CPI refers to the proportion of a given population that would be expected to be included as a potential contributor to an observed DNA mixture. The complement of the CPI is the combined probability of exclusion (CPE). Profile interpretation and CPI calculation involves three steps: assessment of the profile, comparison with reference profiles and inclusion/exclusion determination, and calculation of the statistic.

Prior to comparison with known profiles, peak heights are used to determine whether contributors (i.e., major and minor) can be distinguished. When a known individual's DNA can reasonably be expected to be present, the known contribution can be "subtracted" [13]. When a known cannot be excluded, the calculation is performed for the evidentiary profile irrespective of any known contributor types, etc.).

The advantages of the CPI approach are thought to be its simplicity and the fact that the number of contributors need not be assumed in the calculation. However, even with simplicity, recently, in the U.S., interpretation protocols used for DNA mixtures using the CPI method have been criticized when applied to forensic mixtures for which it is not suited, highlighting issues of effective communication and technology transfer to the end users of the forensic science community [14]. One should be wary of deceptively simple solutions to complex problems as it is possible that the perceived simplicity of the CPI statistic has led in some instances to incorrect applications of the approach. While the number of alleles is used to generate a CPI statistical estimate, it is incumbent upon the user to evaluate a mixture based on the possible genotypes of the contributors and to consider the potential of missing data (i.e., allele drop-out) based on peak height observations at other loci in the profile and the possibility of allele stacking.

If the DNA crime stain profile is low level, then possibility of allele drop-out should be considered. If allele drop out is a reasonable explanation for the observed DNA results, then the CPI statistic cannot be used at those loci in which the phenomenon may have occurred. The formulation of the CPI statistic requires that the two alleles at each locus of the donor being considered must be above the analytical threshold. Hence, if a profile, or a component of it, is low level, additional considerations are needed to ensure that allele drop-out has not occurred at this locus.

While interpretation of a mixture prior to a statistical calculation requires the direct use of peak heights, the assumed number of contributors, the genotype of known contributors or the genotype of persons of interest (POIs), the CPI calculation, in a strict sense, does not require such consideration [13, 15, 16].

The authors recommend moves in favour of using the likelihood ratio (LR) approaches and laboratories have been embracing LR application [17–19]. Use of the LR also must consider the possibility of allele drop-out; but the LR approach has more flexibility than that of the CPI to coherently incorporate the potential for allele drop-out in complex mixtures (i.e., the so-called probabilistic genotyping methods).

If a lab chooses not to convert to using LRs, or if it does intend to convert but is using CPI in the interim, it remains necessary to ensure that when the CPI is used it is applied correctly.

Herein a more explicit description of a DNA mixture protocol is offered with recommendations for applying the CPI. While the approach described herein overall is not a completely new approach to the use of the CPI, it has become essential to formalize the protocol so that proper statistical analyses can be performed when needed in courtroom proceedings. This protocol is provided as one that should be used for applying the CPI when needed.

Calculation of the CPI involves a statistical model that returns an estimate of the sum of the frequencies of all possible genotype combinations included in the observed DNA mixture. While the computation of the statistical estimate, itself, does not require assumptions about the number of contributors, an assumption of the number of contributors is necessary to help inform decisions about whether allele drop-out is likely at particular loci in the evidentiary sample. For example, if only four allelic peaks appear at a locus in a profile assumed to be from two donors, then it is reasonable to assume that allele drop-out has not occurred at that locus.

That there is no published unifying protocol for use of the CPI for evaluation of forensic DNA mixtures has led to some confusion among forensic practitioners on its proper use. Accordingly a detailed protocol is provided herein to guide the community to reduce variation in interpretation and to promote a more defensible application of the CPI. Three publications describe the use of the CPI [13, 20, 21]. All three of these documents correctly recommend that practitioners should not use (i.e., should disqualify) any locus from the CPI calculation that shows, upon evaluation of the DNA results, that allele drop-out is possible. Moreover, all three support the concept that loci that are omitted for calculation of the CPI statistic may still be used for exclusionary purposes.

Given emerging criticism of methods used in forensic DNA mixture analysis, interpretation and statistical evaluation - particularly in the U.S. - it is timely to revisit and reinforce the foundational principles of interpretation of mixtures and subsequent computation as it relates to the CPI (or CPE). The authors recognize and advocate the community as a whole move towards the use of probabilistic genotyping methods [9, 17, 22, 23] with proper validation. However, in the interim, it has become evident that a specific CPI protocol is needed to guide practitioners who currently use it and for reanalysis of past cases in which use of the CPI method may not have considered the guidelines detailed herein. All methods, including probabilistic genotyping and the CPI-based approach, require the ability to deconvolve mixtures.

It is not possible to prescribe rules for every conceivable situation; therefore, it is essential that application of the CPI be performed by well-trained professionals using their judgement and knowledge under the spirit of the guidelines provided herein, their professional education, and relevant experience. Lastly, the protocol described herein is a guideline and does not preclude alternate acceptable methods to interpret DNA mixture evidence as long as the rules applied are always held subservient to the foundational principles involved in proper mixture interpretation.

Methods

Interpretation and application of CPI

Interpretation of a DNA mixture should not be done by simply counting observed alleles. Efforts to deconvolve a mixture into single contributors are advocated where possible [2, 13, 24–26]. If a probative single source profile can be determined at some or all loci then a single-source statistic may be used to calculate a probability estimate (or LR) for that observed profile. This single-source profile may be a deduced major or minor contributor or a deduced foreign contributor by subtracting an assumed known contributor's alleles.

One caution is that single source statistics at some loci and CPI statistics at other loci should never be combined into one statistical calculation [13]. Either use only those loci that enable a single-source deconvolution or the loci that qualify for a CPI calculation. If the two options are investigated, then the statistic with greatest probative value (i.e., the lower probability of the RMP or CPI) should be reported in order to make optimal use of the data available.

Rules for qualifying STR loci for use in CPI/CPE calculations on forensic DNA mixtures

The procedure for DNA mixture interpretation using the CPI approach assumes that a laboratory has an established valid analytical (or detection) threshold (AT), stochastic threshold (ST), stutter filter values (SF), and minimum peak height ratio(s). As PCR is "semi-quantitative" STR allelic peak heights are approximately proportional to the amount of DNA from each donor [2, 24]. One might be able to assume that the peak heights may be equivalent at every locus with very pristine (un-degraded) biological samples, but interpretation should be made on the resultant electropherogram [27, 28]. Typically, across an entire DNA profile, there is a downward trend in peak heights such that longer length PCR amplicons, and therefore the alleles contained within, may exhibit shorter peak heights. This phenomenon is referred to as a "degradation slope" (or "ski slope").

Impact of the number of contributors on DNA mixture interpretation

DNA mixtures involve two or more donors. It is incumbent upon the DNA analyst to carefully assess and state the assumed number of contributors to a profile, even when using the CPI. The SWGDAM STR Interpretation Guidelines [21] 3.4.1. state "For DNA mixtures, the laboratory should establish guidelines for determination of the minimum number of contributors to a sample." While we agree generally, the SWGDAM guidelines are not helpful for the evaluation whether allele drop-out may have occurred. An actual number of contributors, not a minimum number, is needed, as a different number of contributors for the same DNA mixture will result in more or less allele drop-out to explain the observed profile. Consider, for example, a mixture profile with exactly 4 alleles at every locus, under the assumption of a two-person mixture there is no evidence of allele drop-out. However, if the assumption is that there are five contributors for the same mixture profile, then probability of allele drop-out is extremely high.

Each donor may contribute 0, 1, or 2 alleles at each genetic marker (locus) tested (with rare occurrences 3 alleles per locus). Any of the observed peaks (true allelic or backward/forward stutter) may overlap with a peak(s) from the same or another donor of the mixture. When allele or artefact sharing occurs there is an additive effect of the two or more peaks, termed "allele stacking" or "allele masking". As the number of potential contributors increases, so does the uncertainty in accurately determining the true number of contributors [29]. For example, based on the total number of alleles observed across an entire STR profile, it can be extremely difficult, if not impossible, to distinguish a five-person from a six-person DNA mixture and in a number of cases even a three-person from a four-person mixture [29].

These guidelines do not describe in detail how to determine the number of contributors, as a minimum requirement, the number of alleles at each locus and their peak heights should be considered when assigning the number of contributors. Because of the quantity and quality of the DNA being analysed, some loci may meet the determined number of contributors and some may not. For those loci that do not fit the best estimate of the number of contributors, there should be evidence of low signal and/or degradation, which would render the specific locus (or loci) inconclusive for the CPI calculation. Testing additional STR loci may reduce the uncertainty in estimating the potential number of contributors [29]. In addition, challenges arise when close biological relatives have contributed to a mixture or if the DNA is somewhat degraded. Donors to a mixed DNA profile may be referred to as major, minor, and "trace" indicating the relative proportions of their peak heights. For practical purposes minor and "trace" can be considered together as lesser contributors compared with a major contributor(s) of a mixture. In some situations alleles may be missing (i.e., have "dropped out") in evidentiary samples [30–32].

Stutter

Stutter, the inherent by-product of slippage during amplification of STRs, adds complexity to mixture interpretation. Typically, interpretation of whether a peak is solely stutter or stutter along with an allele from another contributor arises when a minor or trace contributor peak(s) is observed at a locus (or other loci) that is similar in height relative to the stutter of the major contributor alleles at the locus. These peaks and their heights are used to help determine whether to qualify or disqualify the locus for use in the CPI calculation.

Stochastic effects

Random variation in peak heights is an inherent property of current DNA typing methodologies. These random variations of peak heights within an individual STR profile or between replicate samples are known as stochastic variation. As the quantity and quality of the input DNA decreases stochastic effects can increase. These effects manifest as variation in peak height between the two peaks at the same locus in a heterozygote or the variation of allele peak heights from the same donor at different loci across the degradation slope line. Such allele peak height variation arises from several factors:

1) Sampling of template from the extract for the aliquot used for the PCR [33],
2) The greater stuttering and lower amplification efficiency of larger alleles (or template accessibility during PCR), and
3) Quality of the template DNA.

It is likely that most of the variation in allele peak heights results from the sampling of template [34, 35] and quality of the sample, but variation during the PCR also contributes, especially with very low template DNA. If the template level is low in the DNA extract then relative variability in the peak heights can be large. This variability is empirically observed and is predicted [36–39]. Because of the strong linear relationship between template (or, more correctly, effective template) and allele peak height, peak height in the actual profile has been a reliable indicator of the presence of stochastic effects and, as such, has been a good indicator for establishing a stochastic threshold (ST) [40, 41].

The ST is the peak height value(s) above which it is reasonable to assume that allele drop-out of a sister allele of a heterozygote has not occurred at a locus [40, 41]. The ST must be determined empirically, based on validation data derived within the laboratory and specific to a given STR kit and analytical instrumentation. Although a binary approach, use of a ST has been deemed important to more formally assess potential allele drop-out. There are several ways in current use to assign a ST (see the Appendix for discussion on setting a ST). A formulaic derivation of the stochastic threshold is displayed in the Additional file 1.

Application of a ST is straightforward for single-source DNA profiles. If a single allele is observed and its peak height is below the ST it is considered possible that a sister allele at that same locus may have dropped out. In contrast to single source samples, in DNA mixtures any given allele peak may actually represent a composite of allele peaks (and depending on position can include stutter peaks). Because of the potential of allele sharing among different contributors to a DNA mixture and the accompanying additive effects in peak heights, a peak height above the ST does not necessarily assure one that a sister allele has not dropped out at that locus. Analysis of the full profile is required to assist in the determination of potential allele drop-out.

Laboratories typically apply a ST for interpretation using a peak height threshold determined based on validation experiments. If the same ST peak height is used across all loci in an entire DNA profile, for many cases involving low level or degraded samples, the loci at the low molecular weight end of the profile (i.e., the smaller amplicons) can exceed the ST whereas at the higher molecular weight end (i.e., the larger amplicons) they may straddle or fall below this threshold.

Role of STR peak heights and PCR amplification stutter artefacts

STR allelic peak heights are approximately proportional to the effective (i.e., amplifiable) amount of DNA from the donor [2, 24]. Typically, across an entire DNA profile, there is a downward trend in peak heights such that longer sized PCR amplicons, and therefore the alleles contained within them, may exhibit shorter peak heights. Such general peak height behavior and locus-specific performance should be considered in DNA mixture interpretation. The possibility of allele dropout at any particular STR locus is assessed, in part, by use of a ST. The phenomenon of allele drop-out was first documented in the early days of PCR-based typing [10, 42]. Indeed, the Scientific Working Group on DNA Analysis Methods (SWGDAM) recognized the use of a ST and stated in [21] Section 3.2.1: “The RFU value above which it is reasonable to assume that, at a given locus, allelic dropout of a sister allele has not occurred constitutes a stochastic threshold.”

Each STR allelic peak may be associated with one backward stutter peak and occasionally a lower signal forward stutter peak [17, 41–44]. At some loci double backward stutter and “N-2” stutter are observed. Therefore, analysts should be familiar with the nuances of each STR marker. In some situations it may be possible for the stutter peaks from one donor to exhibit a similar height to the allelic peaks from another donor. In such instances the potential allele peaks may not be distinguishable from stutter.

Consider a case where it is ambiguous whether a peak is stutter or an allele. In such an instance a contributor with an allele in this ambiguous position would not be excluded. The appropriate inclusion statistic for this locus then includes the allele probabilities for the ambiguous peak positions in the summation for the CPI calculation [13]. Subtraction of the stutter component may assist in determining the signal from the allelic component of that peak. It might be possible to determine that such peaks must be stutter by assuming a certain number of contributors, or a number of minor contributors. For example, if it is reasonable to assume that there is one minor contributor, and two minor allelic peaks already have been identified, then other small peaks in stutter positions can be assumed to represent true stutter.

Results and discussion

Proposed guidelines for an approach to DNA mixture interpretation

The generalized approach is described as follows:

1) Apply a stutter filter as normal and remove any artefacts such as pull-up and spikes.
2) If a single source profile may be deduced from the mixture, then do so. This single-source profile may be a deduced major or minor contributor or a deduced foreign contributor by subtracting an assumed known contributor’s alleles.
   Approaches for calculating single-source statistical estimates of a profile probability can be found in the

National Research Council Report [46]. The random match probability (RMP) describes the estimate of the probability that a randomly selected unrelated person would match the deduced single-source (major or minor) profile from the mixture. If a deduced profile is incomplete at any locus (e.g., one obligate allele, but not the other) is deduced, then this uncertainty should be recorded by some nomenclature such as allele "F" or "any" or some other designator. Often the 2p rule is applied for modified RMP calculations at those specific loci [45, 46]. It is reasonable when interpreting a mixture to "subtract" the profile of any donor who could reasonably be expected (or is assumed) to be present in the sample.

3) If no single-source profile could be deduced or there is some interest in interpreting irresolvable components of the mixture, the CPI approach can be invoked.

To formalize the interpretation the overriding principle (P) for use of loci in CPI calculations is:

P_1: Any locus that has a reasonable probability of allele drop-out should be disqualified from use in calculation of the CPI statistic.

All guidelines that follow are subservient to P_1. Failing to consider the potential of allele drop- out when there are no detectable peaks between the AT and the ST has allowed the often misguided concept to develop that if all observed peaks are above the ST, then the locus unequivocally can be used.

We cannot prescribe what is a "reasonable probability" as the probability relies on the validation performed by the laboratory and on what ST value has been applied (could be overly conservative). However, if a numerical estimation is sought then one could consider allele drop-out no higher than 0.01 being a reasonable value for addressing uncertainty.

With one exception the approach to DNA mixture interpretation should never trump P_1. The exception to P_1 (termed modified or restricted CPI) is an interpretation that can apply to a portion of a profile as opposed to the entire profile. This scenario sometimes occurs where the mixture profile is comprised of multiple major contributors and minor (or trace) contributors where the majors can be resolved readily from the lesser contributing alleles (for example, two major contributors and one minor contributor – (see the section on a major cluster, R_4) [13, 24, 30].

Rule 1 (R_1) locus qualifying rule

A locus is included for use in a CPI calculation if allele drop-out is considered to be highly unlikely. Only qualified loci are used in the calculation of the CPI statistic (Figs. 1 and 2).

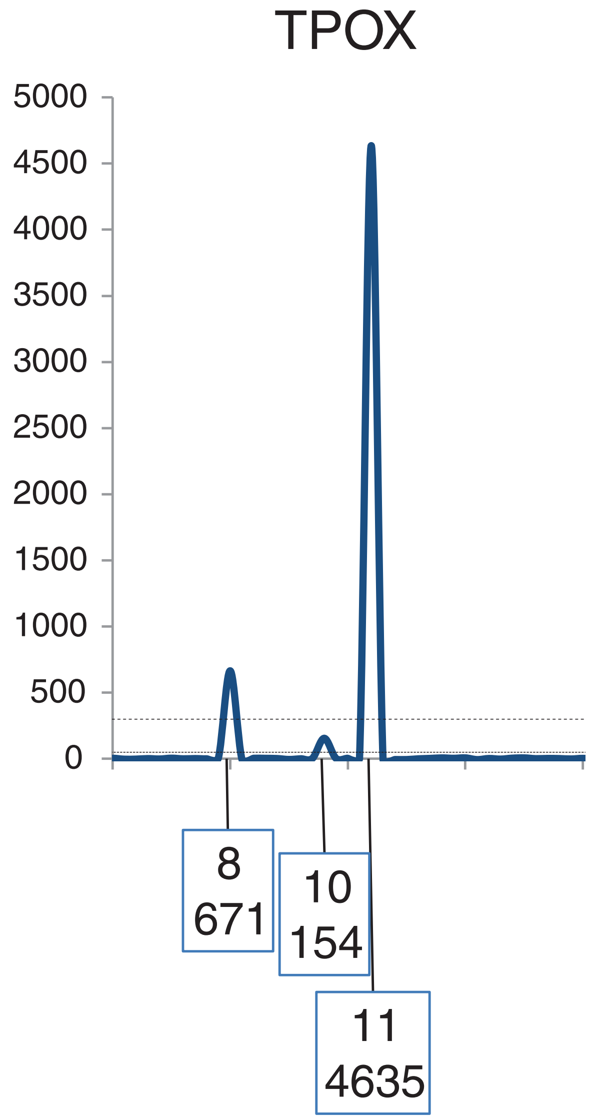


Fig. 1 A depiction of the TPOX locus in an assumed two person mixture. Threshold parameters in this example are: ST = 300 and AT = 50 RFU. If the overall profile supports the best assumption of a two-person mixture, then plausible genotype deconvolution should proceed considering a two-person contribution. The ratio of allele 11:8 is ~7:1. If the contributors donated different amounts to the signal, then plausible genotype deconvolutions to explain the mixture are 8, 8 and 11,11 and 8,11 and 11,11. There is little, if any, possibility of the mixture being derived from an "11,11" and an "8,Q" (where Q stands for an unidentified dropped out allele). Hence, there is no reasonable expectation of allele drop-out, and the locus can be used in the CPI calculation

Guidance (G) for R_1.

$G_{1.1}$: Any locus with an allelic peak height below the ST and above the AT is disqualified for a CPI calculation.

For example, as shown in Fig. 2, this Rule would disqualify loci D3S1358, D16S539, CSF1PO, and TPOX (n.b., under the reinstatement rule described below in

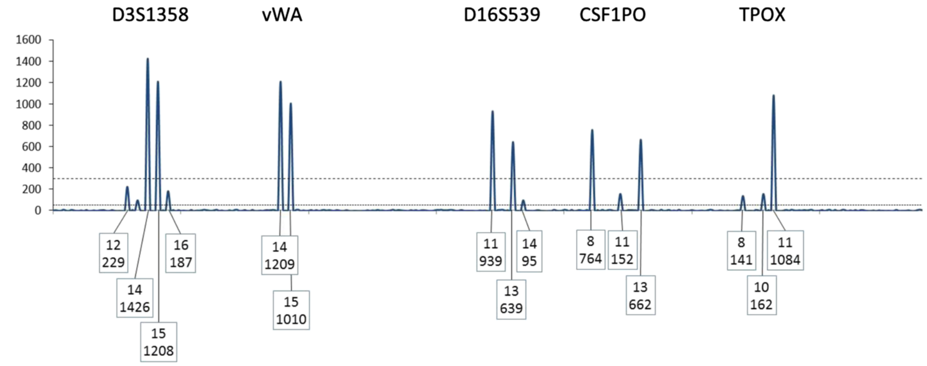


Fig. 2 A depiction of the blue dye channel of a Globalfiler STR profile in an assumed two person mixture. Threshold parameters in this example are: ST = 300 and AT = 50 RFU. At four out of five loci there are visible peaks below the ST that can be assigned as alleles and therefore these four loci are disqualified (Rule 1). At the vWA locus no peaks are observed below the ST. However, allele drop-out is possible, suggesting that the vWA locus also should be disqualified from use in the CPI calculation (note the use of Rule 3 below may allow re-qualification of the D3S1358 locus). N.B., as emphasized in the protocol described herein, a major contributor could be determined readily across the entire profile and if attempted all loci would be interpretable for that purpose

section R_3, it may be possible to re-qualify locus D3S1358).

A locus disqualified for a CPI statistic may still be suitable for an RMP calculation.

$G_{1.2}$: Any locus with an observable peak(s) residing below the AT that is likely to be a true allele(s) is disqualified. A peak below the AT may be deemed to be an allele if there is evidence of low level peaks at other loci, the peak(s) is distinct from the local noise, is not in the "N + 4" (i.e., forward stutter) or "N-4" (i.e., backward stutter) or "N-8" (i.e., −2 repeats) stutter position and has Gaussian morphology. While peaks below the AT are not used for comparison purposes, they might be informative to support the possibility of allele drop-out at the locus (or loci) being evaluated, particularly when there are peaks below the ST (and above the AT) at smaller amplicon loci.

$G_{1.3}$: Evaluation of potential allele drop-out is not constrained to observable peaks at a specific single locus. Instead, a global profile evaluation is required. Any locus that has no allelic peaks below the ST and above the AT but may have an unseen allele(s) (based on the peak heights of alleles at other loci) is disqualified.

Implementation of $G_{1.3}$: If there are minor peaks below or close to the ST or below the AT at other loci, these peaks may be indicators of the potential of allele drop-out. These indicator peaks at other loci should be taken into consideration for potential allele drop-out in the specific locus being evaluated.

R_2: Stutter. Additive effects for alleles overlapping with stutter products must be considered in assessing the potential for allele drop-out at a locus and indistinguishable stutter/allele peaks may need to be included in CPI calculations.

$R_{2.1}$ Check if a peak in a stutter position is considered to have an allele contribution.

$G_{2.1.1}$ To determine whether there is an allele contributing to a peak in the stutter position subtract the stutter threshold or stutter filter value (SF) for the locus from the peak height value for the peak in the stutter position (SPH). The remaining value is the minimum allele contribution (MAC).

SPH − SF = MAC

If MAC > ST, then the locus can be used for use in the CPI calculation.

If MAC ≤ ST, then the locus is disqualified for use in the CPI calculation.

The SF value may not represent the true stutter contribution, as this value often is calculated as the mean stutter + 3SDs. There is a reasonable expectation that the true stutter contribution can be less than the SF value. However, since there is no way to determine the precise stutter contribution, using the maximum value of stutter is advocated.

$G_{2.1.2}$ The locus may be re-qualified (see exception rule R_3 below) even when the MAC ≤ ST, if there is evidence of no allele drop-out at the locus. Evidence of no allele drop-out could come from a deconvolution where all minor or trace alleles have been observed or inferred based on subtraction of an assumed known contributor's alleles. Determining the number of minor contributors (and hence the number of possible minor alleles) can be challenging with complex DNA mixtures. A peak in the stutter position that does not exceed the SF may still have been comprised of both stutter and an allele from another contributor. This peak(s) should be considered potentially allelic based on the data in the profile (Fig. 3).

$R_{2.3}$ If there is no minor allele at this locus but other loci suggest that the height of a possible minor allele at

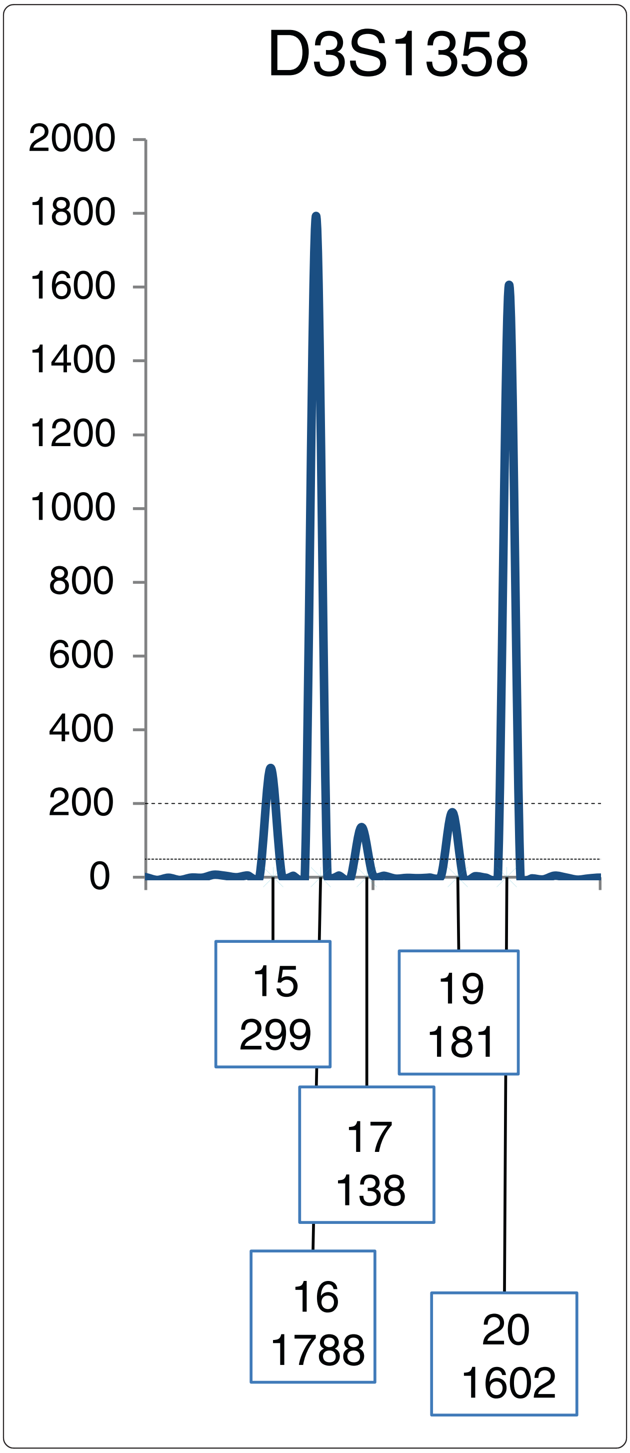


Fig. 3 A depiction of the D3S1358 locus showing a two-person profile. Threshold parameters in this example are: ST = 200 and AT = 50 RFU. Using Identifiler Plus data [47], the stutter filter (SF) for the D3S1358 locus is recommended to be set at 12.27 %. The peak height for allele 16 is 1788 RFUs; thus the stutter threshold for a peak at position 15 is 219 RFUs. The observed peak height at position 15 is 299 RFUs. Therefore, the MAC is 80 RFUs (i.e., 299-219 = MAC). Since 80< ST, the potential for allele drop-out is invoked, and the locus would be disqualified. However, if the overall profile interpretation supports a single minor contributor, then the contributing allele at position 15 can be paired with the minor obligate allele 17 (138 RFUs), and the locus now can be re-qualified (see exception rule R_3), even though both minor allele peak heights are below the ST. While using SWGDAM and ISFG guidelines [18, 19, 21] this example a major profile should be deconvolved, for demonstration purposes a CPI calculation is shown using alleles 15,16,17,20 (the peak at 19 is assumed to represent stutter). $R_{2.2}$ If there is a minor allele of approximately the height of a possible allelic component of a stutter peak and there is at least one minor allele unconfirmed then the stutter peak(s) should be included in the summation for the CPI calculation (Figs. 3 and 4)

this locus is approximately the height of a peak in a stutter position, the stutter peak(s) should be included in the summation for the CPI calculation.

R_3: exception rule. Indicators that alleles below the ST did not drop-out

It is possible to reinstate (requalify) some loci for use in the CPI calculation. This qualification can occur for alleles observed at a locus, dependent on the assumption of the number of contributors to that mixture even where the peak height of an allele(s) falls below the ST (and above the AT). As stated above, while the number of contributors is not taken into account when calculating the CPI, it is imperative that the number of contributors be assumed to determine the potential of allele drop-out. For example, consider a two-person mixture with one major and one minor contributor (Fig. 1), and the assumption of one minor contributor reasonably can be made. If two minor alleles are observed, then the locus may be used in the CPI calculation, regardless of whether any of the minor alleles are below the ST. In this scenario (and other similar ones) there is no indication of allele drop-out at the locus. Referring back to Fig. 2, this qualification would reinstate the D3S1358 locus and allow its use in a CPI calculation.

This approach can be extended to three-person mixtures if interpretation of the overall profile indicates that allele drop-out has not occurred under an assumed number of contributors.

G_3: If a mixture interpretation suggests no drop-out, then the locus can be used in the CPI calculation.

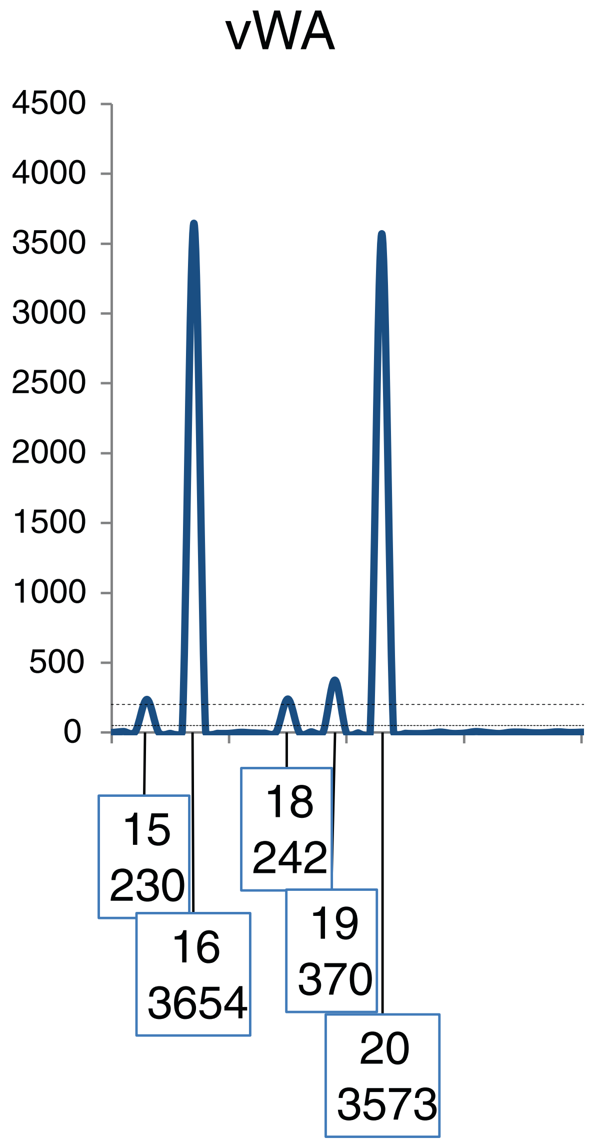


Fig. 4 A depiction of the vWA locus illustrating the application of R_2. Threshold parameters in this example are: ST = 200 and AT = 50 RFU. Hence the obligate minor allele at 18 is above the ST and drop-out of its sister allele is unlikely. This locus is qualified for use in the CPI calculation. Under the assumption of two contributors there is one minor allele unconfirmed. Both the 15 and 19 peaks are below the stutter filter (SF) and hence could be all stutter or a composite of stutter and allele. This example illustrates the scenario where peaks in the stutter position fall below the SF. The partner to the 18 allele must be at one of the positions 15,16,18,19, or 20. Since the minor contributor genotype cannot be resolved with sufficient confidence, for this example the probability of inclusion is calculated as $PI = (p_{15} + p_{16} + p_{18} + p_{19} + p_{20})^2$

$G_{3.1}$: If all possible alleles are observed (e.g., a two-person mixture and 4 alleles), then the locus can be used in the CPI calculation.

R_4: major cluster rule

If a set of peaks representing more than one donor is distinct from one or more minor or trace peaks then the CPI approach may be applied to the “major cluster” (see $G_{4.1}$, Fig. 5, Table 1). We outline an algorithm to confirm a major cluster (see Appendix).

$G_{4.1}$: To qualify a locus for use with a major cluster, first there must be a clear visual distinction between a set of large peaks and a set of trace peaks. The principle is that all major peaks must be identifiable and for these major peaks allele drop-out must be deemed unlikely.

There are two aspects to this principle;

$G_{4.1.1}$ Any allele peak assigned to the major cluster
must be sufficiently high that it could not have a
partner allele in the minor set, and

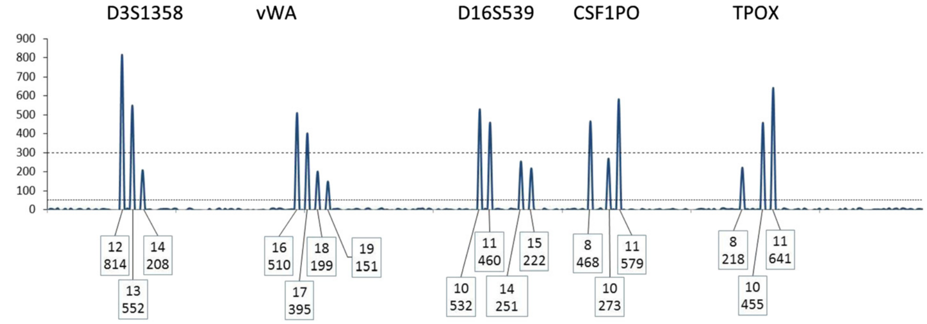


Fig. 5 A depiction of a hypothetical depiction of the blue dye channel of a Globalfiler electropherogram in an assumed two person mixture. Threshold parameters in this example are: ST = 300 and AT = 50 RFU. A POI who is 13,13 at D3S1358 would support an exclusion with two assumed contributors. If this POI were included then the other contributor would have to be 12, 14 at the locus with an improbable PHR

$G_{4.1.2}$ Allele peaks assigned to the major cluster must be sufficiently high that allele drop-out is unlikely even when consideration is given that the peak might be a composite of major and minor.

$G_{4.2}$: This assessment requires some level of deconvolution and is more straightforward if there are only two major profiles and one trace contributor. Consider a locus with four large peaks and two small ones (Fig. 6). Such a profile (at this single locus) is consistent with being from two major profiles and one trace profile. In such a case determine that a trace peak and a major peak cannot be misassigned. If there are only three, two, or one major peaks present, check that any peaks assigned as trace could not be a major peak. This approach is best accomplished by visualizing the major and trace peaks across the entire profile and fitting realistic degradation curves. If there is no distinction between a set of large peaks and the small ones at a locus (or loci), then assigning a "major cluster" should not be attempted (Figs. 7 and 8).

R_5. calculation of CPI/CPE

The formula for calculating the CPI has been described elsewhere [10] (Appendix). For each of the qualifying loci sum the allele frequencies for the allelic or potentially allelic peaks (those peaks added by the stutter rule) and square that value. Multiply the value of each locus that qualified under the assumption of independence to produce the CPI (n.b., the CPE is 1-CPI).

Table 1 The peak height analysis for the STR profile shown in Fig. 6

Locus	*SMP*	*LTP*	$\frac{LTP(2NT-T)}{PHRL}$	Pass/fail	$SMP-\frac{LTP(2NT-T)}{PHRL}$	Major cluster
D3S1358	1698	290	580	Pass	1118	Qualified
vWA	1648	289	578	Pass	1070	Qualified
D16S539	1386	336[a]	672	Pass	714	Qualified
CSF1PO	1380	206	412	Pass	968	Qualified
TPOX	1727	289	578	Pass	1149	Qualified

SMP smallest main peak, *LTP* largest trace peak, *NT* number of trace contributors, *T* number of trace alleles, *PHRL* peak height ratio limit value.
[a]The "9" peak at D16S539 may be larger because of a stutter component. Hence LTP is 336 RFU or less

$G_{5.1}$ With the exception of using data from a reference profile in which an assumption of one of the contributors is known, such as from an "intimate" sample (described in G5.3), comparison of a DNA mixture profile with that of a known suspect/victim or other known POIs, when possible, should not be carried out until the mixture evidence has been fully evaluated as described above. Comparison of the evidence and known profiles for inclusion/exclusion purposes is independent of the CPI statistical calculation. Regardless, All the alleles of the POIs should have a corresponding allelic or potentially allelic peak in the qualifying loci. If the evidence supports an exclusion, the calculation of the CPI is unnecessary for that comparison. If there is a failure to exclude based on genotype possibilities derived from peak heights at qualified or disqualified loci, then a computation is provided. Computation of the CPI does not require examination of the STR profile (genotypes) of the known individuals (suspect, victim, POIs). At the point of computation of the CPI, the DNA mixture profile is composed of qualified and disqualified loci.

$G_{5.2}$ There can be only one value for the CPI/CPE computed for each DNA mixture profile. The interpretation of potential allele drop-out should be made prior, when possible, to evaluating known reference samples. Adjustments to fit the interpretation to reject or reinstate a locus based on additional information from a person of interest profile (i.e., confirmation bias and fitting the profile interpretation to explain missing data based on a known sample) are inappropriate [48–50].

$G_{5.3}$ One exception to using data from a reference profile is where an assumption of one of the contributors is known, such as from an "intimate" sample. The assumption of the individual(s) being a known contributor(s) must be documented. In situations where a contributor(s)

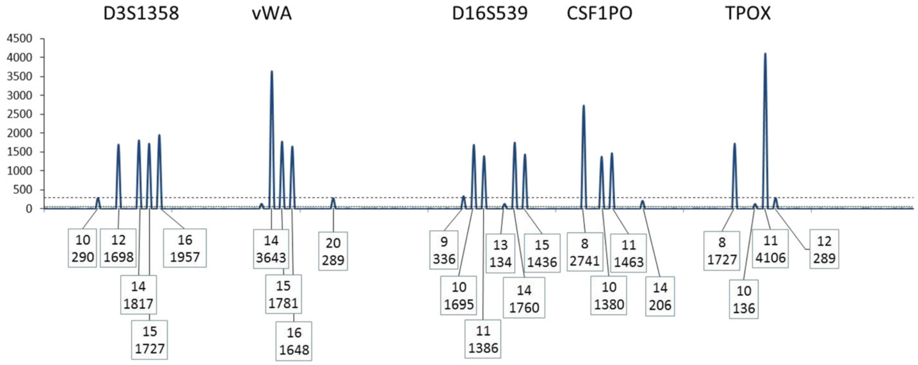


Fig. 6 A depiction of the blue dye channel of a Globalfiler electropherogram. Threshold parameters in this example are: ST = 200 and AT = 50 RFU. This example is an acceptable "major cluster". There is one trace contributor (NT = 1). For this example a peak height ratio limit (PHRL) of 0.50 is used (See Table 1 for peak height analysis using the major cluster rules). The PHRL should be determined by each laboratory based on validation studies

is assumed, then subtraction of the alleles of the known contributor(s) is justified (which also may be applied to major cluster analyses).

$G_{5.4}$ Use of the 2p rule for the CPI is not valid.

The 2p statistic was designed for single-source samples where one allele was present at a locus and there was strong support for allele drop-out.

Exculpatory evidence

Once the mixture has been evaluated, both the qualified and disqualified loci should be inspected for potential exclusionary evidence. For the qualified loci exclusionary evidence may be based on the absence of alleles or the absence of deconvolved genotypes in the mixture compared with those of the known reference profile. If the deconvolved genotypes of the mixture are different from the genotype of the known comparison profiles, then an exclusion interpretation is supported. If the locus (or loci) was deemed disqualified for the CPI calculation, allele drop-out should be considered when including or excluding a potential donor.

R_6. For the qualified loci exclusionary evidence exists when the POI has any alleles not present in the crime stain profile.

Consider the D3S1358 locus shown in Fig. 2. The rest of the profile supports a two-person mixture. Initially this locus would be disqualified based on having peaks below the ST but then is reinstated because both minor peaks are present. At this locus a POI is excluded if the POI has any allele outside the set [12, 13, 15, 16].

R_7. For loci that can be deconvolved exculpatory evidence exists when the POI has a genotype not amongst the set of supported genotypes.

Consider again the D3S1358 locus (Fig. 2). At this locus a POI is excluded if the POI has a genotype other than the genotypes {12,16 or 14,15}.

R_8. For disqualified loci exculpatory evidence can occur but relies on the profile, allowing for missing data, to determine if the POI is unlikely to be a donor.

G_8. The POI is unlikely to be a donor if the allele(s) consistent with the POI and the total number of observed alleles at a given locus invalidate or do not support the assumed number of contributors to the DNA mixture. The

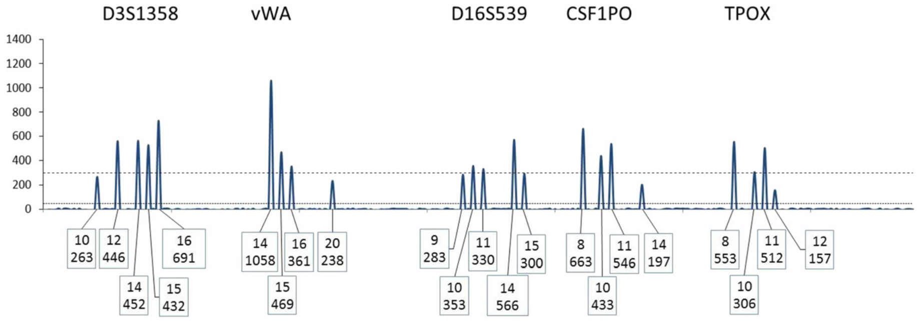


Fig. 7 A depiction of the blue dye channel of a Globalfiler STR profile. Threshold parameters in this example are: ST = 300 and AT = 50 RFU. This example is an unacceptable major cluster. There is one minor contributor (NT = 1). The two major profiles are not much greater in height than the minor profile

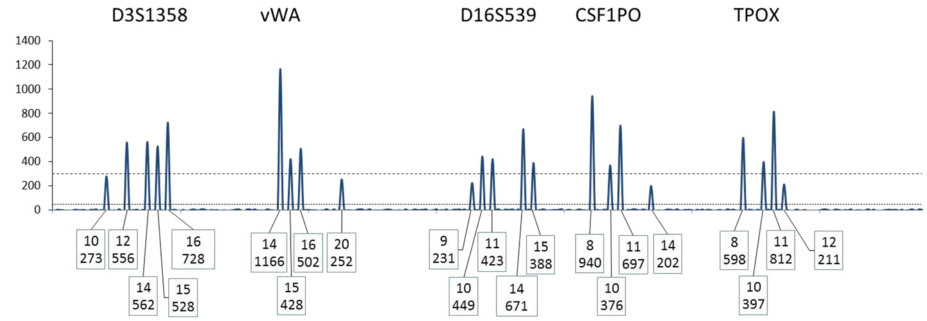


Fig. 8 A depiction of the blue dye channel of a Globalfiler electropherogram. Threshold parameters in this example are: ST = 300 and AT = 50 RFU. This example is intended to illustrate an unacceptable major cluster. There is one minor contributor (NT = 1). The two major profiles look to be about twice the height of the minor. A PHRL of 0.50 is used for this example. See Table 2 for the peak height analysis using the major cluster rule

inclusion of the POI would cause a mismatch with the assumed number of contributors (Fig. 5). Before finalizing an exclusion ensure that the assumed number of contributors holds throughout the profile. If that assumption is not valid, the result may be considered inconclusive.

Conclusions

The path forward

The protocol described herein is intended to help reduce confusion and misunderstanding in the forensic community about how to best apply the CPI in evaluation of forensic DNA mixtures, not only for current casework but for retrospective review of past cases. While the protocol detailed herein is not novel in the sense that most aspects of the CPI have been discussed in the literature, the lack of a unifying detailed CPI protocol has led to confusion and in some cases misapplication of this method. For this reason it is important that a detailed DNA mixture interpretation protocol be offered to reduce inter- and intra-laboratory variation in application of the CPI. Cases for which a CPI was calculated without considering the possible presence of allele drop-out or other stochastic effects might benefit from a thorough scientific review. Other cases for review could include those in which multiple CPIs were computed on the same mixture profile, or when confirmation bias was possible (e.g., when "suspect-driven" mixture analysis was performed).

In Texas, the Forensic Science Commission has been working with laboratories to assess the DNA mixture protocols and review the statistical analyses in selected cases using the CPI/CPE. For laboratories or jurisdictions that modify their DNA mixture interpretation protocols, either in light of this document or for other reasons, there may be reason to review a sample of selected pending or previously reported DNA mixture casework. Forensic laboratories can work closely with all stakeholders in their respective jurisdictions to address these issues in a collaborative and constructive manner.

Appendix

Determination and use of stochastic thresholds

Several approaches have been used to determine stochastic thresholds. These include:

1) Methods based on largest surviving allele,
2) Methods based on peak height ratio studies, and
3) Methods based on assigning a probability of drop-out, Pr (D).

We have not specifically tested these different methods against each other and hence do not recommend one method over the other. There is a compromise required when setting the ST: The higher that it is set the lower the risk that dropout is actually possible, but the more information that is wasted.

The ST must be empirically determined based on data derived within the laboratory and specific to a given amplification kit and the detection instrumentation used? The laboratory should evaluate the applicability of the ST among multiple instruments (i.e., is one CE more sensitive than others?). If measures are used to enhance detection sensitivity (e.g., increased amplification cycle number, increased injection time), the laboratory should perform additional studies to establish a separate stochastic threshold(s).

1. Methods based on the largest surviving allele

In this method a study is made of DNA samples constructed from known donors so that the genotypes of the input DNA are known with certainty (known ground truth). Often these samples are pristine and single source. Input DNA amounts that span the range over which allele "drop-out" is expected are amplified. Loci where the known ground truth is a heterozygote profile and where one allele has dropped out are noted and the height of the surviving allele is recorded. The ST is placed at some position with regard to the height of the "surviving allele". This placement often is the maximum peak height observed.

While useful, the surviving allele method does not directly address the probability of allele drop-out at the ST. This method of determining the ST is limited by sampling, such that the larger the number of samples, the greater is the chance of observing allele drop with a higher surviving partner peak height. Consider that a value for ST is chosen from a dataset of size N such that allele drop-out has never been observed with a surviving allele higher than this threshold. It is tempting to conclude that at this ST the probability of allele drop-out is zero. In reality, if N is deemed large (e.g., 100 different profiles in the stochastic range), then the probability of allele drop-out at this ST will be small but likely not zero. In contrast, if a much larger sample (e.g., $N = 1000$) is used there will be a possibility of some surviving alleles with dropped allelic ("sister allele") partners above the previous ST.

2. Methods based on peak height ratio studies

In this method of implementing a ST, the same type of data as described above can be used. It is valuable to analyze down to below the AT that will be used in casework as this analysis helps with the average peak height (APH) for low level data. For example, if it is proposed to use 50 RFUs as an AT in forensic casework, then it may be suitable to analyze samples down to as low as 20 or 25 RFUs (the "research AT"). The peak height ratio (PHR) and the APH for each heterozygote locus then is determined. Missing data (alleles that have dropped out below the research AT) are input at some value (e.g., half the research AT) to determine the APH and PHR. A plot then is made of PHR vs APH. A curve (the peak height ratio limit, *PHRL*) is fitted to these data of the form $PHRL = \frac{k}{\sqrt{APH}}$ that captures all or most of the data. This value should be set to capture 0.995 of the data. This setting can be done simply by plotting the line $PHRL = \frac{k}{\sqrt{APH}}$ on the graph of *PHR* v *APH* and varying the value for *k*. Once *k* is assigned then $\log \frac{ST}{AT} = \frac{k}{\sqrt{\frac{AT+ST}{2}}}$. This equation has no algebraic solution and has to be solved by numerical means. The probability of dropout when the surviving peak is at the ST is approximately the fraction of the data not captured by the fitted curve (using the recommended value of 0.995 this is 0.005). This method can explicitly obtain the probability of allele drop-out.

3. Methods based on assigning a probability of drop-out, Pr (D)

In this method of ST placement the same type of data as described above can be used. As stated above, it is valuable to analyze down to below the AT that will be used in forensic casework. The method described in [33] is used to calculate a function giving the probability of allele drop-out and produces constants 0 and 1. If α is the probability of allele drop-out accepted by the laboratory for the ST (e.g., 1 in 1000) then $\mathrm{ST} = e^{\frac{\ln\left(\frac{\alpha}{1-\alpha}\right)-\beta_0}{\beta_1}}$ where β_0 and β_1 are coefficients from the logistical regression. Timken and colleagues discuss use of a closely similar approach [31].

If the ST is applied as is typically done (i.e., an allele above the ST is assumed to have a partner that has not dropped out), then the probability of allele drop-out is technically larger by an unknown amount. This expectation is because the probability of allele drop-out is assigned from the expected height of peaks at this locus based on the entire profile across all loci, and not simply the height of one allele peak.

CPI/CPEThe inclusion probability also can be defined as: the probability that a random person would be included as a contributor to the observed DNA mixture. The complement of the *CPI* is the combined probability of exclusion (CPE). It proceeds in two steps, an inclusion/exclusion phase followed by the calculation of a statistic. When a person of interest is not excluded then: If the mixture has alleles $A_1 \ldots A_n$ then the inclusion probability at locus *l*, (PI_l) is $PI_l = \left(\sum_i p(A_i)\right)^2$ if Hardy-Weinberg Equilibrium expectations are assumed. By writing $\sum_i p(A_i) = p$ the $PI_l = p^2$ is obtained.

The *PI* across multiple loci (*CPI*) is calculated as $CPI = \prod_l PI_l$

A suggested algorithm for confirming a major cluster

The following algorithm is based on a valid peak height ratio limit value, PHRL. Determine the largest trace peak (*LTP*) and number of the minor or trace contributor(s) at that locus (*NT*). This evaluation should be done by considering all minor or trace peaks at this locus along with any indicator peaks (trace alleles) at other loci. Apply a plausible degradation curve to the profile if needed. Check that LTP is not low with regard to the trace peaks at other loci. If it is, one can adjust its height upwards.

One way to qualify a locus for use as a "major cluster" is to consider the smallest major peak (SMP):

The sum of the heights of all unseen trace peaks (2NT-T), where T is the number of trace peaks observed, is not expected to exceed the value computed by $\frac{LTP(2NT-T)}{PHRL}$.

If $SMP > \frac{LTP(2NT-T)}{PHRL}$ then this peak must have a component from a major contributor in it.

Table 2 The peak height analysis using the major cluster rule for the STR profile shown in Fig. 8. A visual inspection alone should suggest that a major cluster cannot be assigned for this profile since there is no clear separation between a set of large peaks and smaller ones. However the analysis is performed for demonstration purposes

Locus	Major cluster peaks	SMP	*LTP*	$\frac{LTP(2NT-T)}{PHRL}$ or $\frac{LTP}{PHRL}$	Pass/ fail	$SMP-\frac{LTP(2NT-T)}{PHRL}$
D3S1358	Top four	528	273	546	Fail	All terms in this column are negative.
vWA	Top three	428	252	504	Fail	
D16S539	Top Four	388	231	462	Fail	
CSF1PO	Top three	376	202	404	Fail	
TPOX	Top three	397	211	422	Fail	

Check that this component is large enough that allele drop-out is unlikely. This assumption of no allele drop-out is expected to be true if the smallest major component exceeds the *ST*. Test this assumption by application of the inequality $SMP-\frac{LTP(2NT-T)}{PHRL} > ST$ otherwise the locus is disqualified.

If $T = 2\text{NT}$ then $SMP > \frac{LTP}{PHRL}$

If the *SMP* is small (e.g., less than *ST*) it is likely that the PHR is too large and the formulas cannot be relied upon (Figs. 6 and 7, Table 2). While these specific rules have not been described in detail (although inferred in [12]) they may appear novel. However, they derive deductively from the PHR. The validity of this rule relies on the validation of the laboratory's PHR.

Additional file

Additional file 1: A Supplemental Materials section is provided which shows a formulaic derivation of the stochastic threshold. (DOC 251 kb)

Acknowledgements
The authors thank Todd Bille (BATFE Laboratory, Beltsville, MD), Christina Capt (University of North Texas Health Science Center, Ft Worth, TX), Lynn Garcia (Texas Forensic Science Commission, Austin, TX), Carll Ladd (CT Forensic Science Laboratory, Meriden, CT), Roland van Oorschot (La Trobe University, Melbourne, Australia), and Bruce Weir (University of Washington, Seattle, WA) for helpful discussions. Points of view in this document are those of the authors and are not made to represent any of the institutions with which the authors are affiliated and do not necessarily represent the official position or policies of the U.S. Departments of Justice or Commerce.

Funding
Supported, in part, by grant 2014-DN-BX-K028 from the US National Institute of Justice.

Availability of data and materials
All supporting data are included in manuscript.

Authors' contributions
FRB, JSB, and BB were they main architects of the manuscript. All authors (FRB, JSB, BB, JMB, MDC) contributed to the preparation of the original draft of the manuscript including all subsequent edits and revisions prior to submission. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests. John Buckleton is one of the developers of STRmix^{TM} but does not receive any benefit, direct or indirect, from sales of this software product.

Consent for publication
All co-authors hereby consent to publish.

Ethical approval and consent to participate
Not applicable.

Individual persons data
Not applicable.

Author details
[1]Center for Advanced Molecular Diagnostics, Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, 75 Francis Street, Boston, MA 02115, USA. [2]ESR (The Institute of Environmental Science and Research), Private Bag 92021, Auckland 1142, New Zealand. [3]Statistical Engineering Division, National Institute of Standards and Technology, 100 Bureau Drive, Mail Stop 8980, Gaithersburg, MD 20899, USA. [4]Department of Molecular and Medical Genetics, Institute of Applied Genetics, University of North Texas Health Science Center, 3500 Camp Bowie Blvd., Fort Worth, TX 76107, USA. [5]National Institute of Standards and Technology, Special Programs Office, 100 Bureau Drive, Mail Stop 4701, Gaithersburg, MD 20899, USA. [6]National Institute of Standards and Technology, Applied Genetics Group, 100 Bureau Drive, Mail Stop 8314, Gaithersburg, MD 20899, USA.

Received: 5 August 2016 Accepted: 18 August 2016
Published online: 31 August 2016

References

1. Torres Y, Flores I, Prieto V, Lopez-Soto M, Farfan MJ, Carracedo A, et al. DNA mixtures in forensic casework: a 4-year retrospective study. Forensic Sci Int. 2003;134:180–6.
2. Clayton T, Whitaker JP, Sparkes RL, Gill P. Analysis and interpretation of mixed forensic stains using DNA STR profiling. Forensic Sci Int. 1998;91:55–70.
3. Ladd C, Lee HC, Yang N, Bieber FR. Interpretation of complex forensic DNA mixtures. Croat Med J. 2001;42:244–6.
4. Holt CL, Buoncristiani M, Wallin JM, Nguyen T, Lazaruk KD, Walsh PS. TWGDAM validation of AmpFlSTR PCR amplification kits for forensic DNA casework. J Forensic Sci. 2002;47:66–96.
5. LaFountain MJ, Schwartz MB, Svete PA, Walkinshaw MA, Buel E. TWGDAM validation of the AmpF/STR Profiler Plus and AmpF/STR COfiler STR multiplex systems using capillary electrophoresis. J Forensic Sci. 2001;46:1191–8.
6. Wallin JM, Buoncristiani MR, Lazaruk KD, Fildes N, Holt CL, Walsh PS. TWGDAM validation of the AmpFlSTR Blue PCR Amplification Kit for forensic casework analysis. J Forensic Sci. 1998;43:854–70.
7. Wilson MR, Polanskey D, Butleer J, DiZinno JA, Replogle J, Budowle B. Extraction, PCR amplification, and sequencing of mitochondrial DNA from human hair shafts. Biotechniques. 1995;18:662–9.
8. Bornman DM, Hester ME, Schuetter JM, Kasoji MD, Minard-Smith A, Barden CA, Nelson SC, Godbold G, Baker C, Yang B, Walther JE, Tornes IE, Yan PS, Rodriguez B, Bundschuh R, Dickens ML, Young BA, Faith SA. Short-read, high-throughput sequencing technology for STR genotyping. Biotech Rapid Dispatches. 2012;2012:1–6.
9. Gill P, Whitaker JP, Flaxman C, Brown N, Buckleton JS. An investigation of the rigor of interpretation rules for STR's derived from less that 100 pg of DNA. Forensic Sci Int. 2000;112:17–40.
10. Walsh PS, Erlich HA, Higuchi R. Preferential PCR amplification of alleles: mechanisms and solutions. PCR Methods Appl. 1992;1:241–50.
11. Devlin B. Forensic inference from genetic markers. Stat Methods Med Res. 1993;2:241–62.
12. DNA Advisory Board. Statistical and population genetics issues affecting the evaluation of the frequency of occurrence of DNA profiles calculated from pertinent population database(s). Forensic Science Communications. 2000;2(3):1–8.
13. Budowle B, Onorato AJ, Callaghan TF, Della Manna A, Gross AM, Guerreri RA, et al. Mixture Interpretation: defining the relevant features for guidelines for the assessment of mixed DNA profiles in forensic casework. J Forensic Sci. 2009;54:810–21.

14. Committee on Identifying the Needs of the Forensic Sciences Community, National Research Council. Strengthening Forensic Science in the United States: A Path Forward. Document No.:228091. 2009.
15. Buckleton J, Curran J. A discussion of the merits of random man not excluded and likelihood ratios. Forensic Sci Int Genet. 2008;2:343–8.
16. Curran JM, Buckleton JS. Inclusion Probabilities and Dropout. J Forensic Sci. 2010;55:1171–3.
17. Balding DJ, Buckleton J. Interpreting low template DNA profiles. Forensic Sci Int Genet. 2009;4:1–10.
18. SWGDAM, Guidelines for the Validation of Probabilistic Genotyping Systems http://media.wix.com/ugd/4344b0_22776006b67c4a32a5ffc04fe3b56515.pdf. Accessed 22 Aug 2016.
19. Gill P, Brenner CH, Buckleton JS, Carracedo A, Krawczak M, Mayr WR, et al. DNA commission of the International Society of Forensic Genetics: recommendations on the interpretation of mixtures. Forensic Sci Int. 2006;160:90–101.
20. Butler JM. Advanced Topics in Forensic DNA Typing: Interpretation. Oxford: Elsevier; 2015.
21. Scientific Working Group on DNA Analysis Methods (SWGDAM). SWGDAM Interpretation Guidelines for Autosomal STR Typing by Forensic DNA Testing Laboratories. 2010. Available at www.swgdam.org. Accessed 22 Aug 2016.
22. Perlin MW, Legler MM, Spencer CE, Smith JL, Allan WP, Belrose JL, et al. Validating TrueAllele® DNA mixture interpretation. J Forensic Sci. 2011;56:1430–47.
23. Taylor D, Bright J-A, Buckleton J. The interpretation of single source and mixed DNA profiles. Forensic Sci Int Genet. 2013;7:516–28.
24. Gill P, Sparkes RL, Pinchin R, Clayton T, Whitaker JP, Buckleton JS. Interpreting simple STR mixtures using allelic peak areas. Forensic Sci Int. 1998;91:41–53.
25. Bill M, Gill P, Curran J, Clayton T, Pinchin R, Healy M, et al. PENDULUM - A guideline based approach to the interpretation of STR mixtures. Forensic Sci Int. 2005;148:181–9.
26. Clayton TM, Buckleton JS. Mixtures. Forensic DNA Evidence Interpretation. Boca Raton: CRC Press; 2004. p. 217–74.
27. Bright J-A, Taylor D, Curran JM, Buckleton JS. Degradation of forensic DNA profiles. Aust J Forensic Sci. 2013;45:445–9.
28. Tvedebrink T, Eriksen PS, Mogensen HS, Morling N. Statistical model for degraded DNA samples and adjusted probabilities for allelic drop-out. Forensic Sci Int Genet. 2012;6:97–101.
29. Coble MD, Bright J-A, John B, Curran JM. Uncertainty in the number of contributors in the proposed new CODIS set. Forensic Sci Int Genet. 2015;19:207–11.
30. Gill P, Haned H, Bleka O, Hansson O, Dorum G, Egeland T. Genotyping and interpretation of STR-DNA: Low-template, mixtures and database matches - twenty years of research anddevelopment. Forensic Sci Int Genet. 2015;18:100–17.
31. Steele CD, Balding DJ. Statistical evaluation of forensic DNA profile evidence. Annu Rev Stat Appl. 2014;1:1–20.
32. Steele CD, Greenhalgh M, Balding DJ. Verifying likelihoods for low template DNA profiles using multiple replicates. Forensic Sci Int Genet. 2014;13:82–9.
33. Ge J, Budowle B. Modeling one complete versus triplicate analyses in low template DNA typing. Int J Legal Med. 2014;128:259. doi:10.1007/s00414-013-0924-6 and Erratum Int J Legal Med (2014). 128:733. doi:10.1007/s00414-014-0992-2.
34. Gill P, Curran J, Elliot K. A graphical simulation model of the entire DNA process associated with the analysis of short tandem repeat loci. Nucleic Acids Res. 2005;33:632–43.
35. Timken MD, Klein SB, Buoncristiani MR. Stochastic sampling effects in STR typing: Implications for analysis and interpretation. Forensic Sci Int Genet. 2014;11:195–204.
36. Weusten J, Herbergs J. A stochastic model of the processes in PCR based amplification of STR DNA in forensic applications. Forensic Sci Int Genet. 2012;6:17–25.
37. Bright J-A, McManus K, Harbison S, Gill P, Buckleton J. A comparison of stochastic variation in mixed and unmixed casework and synthetic samples. Forensic Sci Int Genet. 2012;6:180–4.
38. Bright J-A, Huizing E, Melia L, Buckleton J. Determination of the variables affecting mixed MiniFiler(TM) DNA profiles. Forensic Sci Int Genet. 2011;5:381–5.
39. Bright J-A, Turkington J, Buckleton J. Examination of the variability in mixed DNA profile parameters for the Identifiler(TM) multiplex. Forensic Sci Int Genet. 2009;4:111–4.
40. Moretti T, Baumstark AL, Defenbaugh BS, Keys KM, Smerick JB, Budowle B. Validation of short tandem repeats (STRs) for forensic usage: performance testing of fluorescent multiplex STR systems and analysis of authentic and simulated forensic samples. J Forensic Sci. 2001;46:647–60.
41. Moretti TR, Baumstark AL, Defenbaugh DA, Keys KM, Brown AL, Budowle B. Validation of STR typing by capillary electrophoresis. J Forensic Sci. 2001;46:661–76.
42. Budowle B, Lindsey JA, DeCou JA, Koons BW, Giusti AM, Comey CT. Validation and population studies of the loci LDLR, GYPA, HBGG, D7S8, and Gc (PM loci) and HLA-DQ-alpha using a multiplex amplification and typing procedure. J Forensic Sci. 1995;40:45–54.
43. Ensenberger MG, Thompson J, Hill B, Homick K, Kearney V, Mayntz-Press KA, et al. Developmental validation of the PowerPlex® 16 HS System: An improved 16-locus fluorescent STR multiplex. Forensic Sci Int Genet. 2010;4:257–64.
44. Laurin N, DeMoors A, Frégeau C. Performance of Identifiler Direct and PowerPlex 16 HS on the Applied Biosystems 3730 DNA Analyzer for processing biological samples archived on FTA cards. Forensic Sci Int Genet. 2012;6:621–9.
45. Budowle B, Guisti AM, Waye JS, Baechtel FS, Fourney RM, Adams DE, et al. Fixed bin analysis for statistical evaluation of continuous distributions of allelic data from VNTR loci. Am J Hum Genet. 1991;48:841–55.
46. National Research Council (US) Committee on DNA Forensic Science. The Evaluation of Forensic DNA Evidence: an update. Washington, DC: National Academies Press (US); 1996.
47. AmpFISTR® Identifiler®Plus PCR Amplification Kit User Guide Publication Number 4440211. Foster: Applied Biosystems, Life Technologies; 2015. https://www3.appliedbiosystems.com/cms/groups/applied_markets_marketing/documents/generaldocuments/cms_076395.pdf
48. Dror IE, Charlton D, Peron AE. Contextual information renders experts vulnerable to making erroneous identifications. Forensic Sci Int. 2006;156:74–8.
49. Bille TW, Bright JA, Buckleton JS. Application of random match probability calculations to mixed STR profiles. J Forensic Sci. 2013;52:474–85.
50. Dror IE, Hampikian G. Subjectivity and bias in forensic DNA mixture interpretation. Sci Justice. 2011;51(4):204–8.

Exhibit 5

Hinda Haned, et al., Validation of probabilistic genotyping software for use in forensic DNA casework: Definitions and illustrations

Science and Justice 56 (2016) 104–108

Contents lists available at ScienceDirect

Science and Justice

journal homepage: www.elsevier.com/locate/scijus

Validation of probabilistic genotyping software for use in forensic DNA casework: Definitions and illustrations

Hinda Haned [a,*], Peter Gill [b,c], Kirk Lohmueller [d], Keith Inman [e], Norah Rudin [f]

[a] Netherlands Forensic Institute, Department of Human Biological traces, The Hague, The Netherlands
[b] Norwegian institute of Public Health, Oslo, Norway
[c] Department of Forensic Medicine, University of Oslo, Norway
[d] Department of Ecology and Evolutionary Biology, University of California, Los Angeles, 621 Charles E. Young Drive South, Los Angeles, CA 90095-1606, United States
[e] Department of Criminal Justice Administration, California State University, East Bay, 4069 Meiklejohn Hall, 25800 Carlos Bee Boulevard, Hayward, CA 94542, United State
[f] 650 Castro Street, Suite 120-404, Mountain View, CA 94041, United States

ARTICLE INFO

Article history:
Received 18 June 2015
Received in revised form 26 November 2015
Accepted 27 November 2015

Keywords:
Validation
Likelihood ratio
Probabilistic
Genotyping
Strength of evidence
LRmix

ABSTRACT

A number of new computer programs have recently been developed to facilitate the interpretation and statistical weighting of complex DNA profiles in forensic casework. Acceptance of such software in the user community, and subsequent acceptance by the court, relies heavily upon their validation. To date, few guidelines exist that describe the appropriate and sufficient validation of such software used in forensic DNA casework. In this paper, we discuss general principles of software validation and how they could be applied to the interpretation software now being introduced into the forensic community. Importantly, we clarify the relationship between a statistical model and its implementation via software. We use the LRmix program to provide specific examples of how these principles can be implemented.

1. Background and scope

A number of new computer programs have recently been developed to facilitate the interpretation and statistical weighting of complex DNA profiles in forensic casework (for a review, see [1]). Complex profiles may encompass a multitude of confounding factors resulting from DNA profiling of a low quantity and/or low quality biological sample. The resulting profile may contain multiple contributors, may lack information from the true contributors (allelic drop-out), may include extraneous information unrelated to the crime-sample information (allelic drop-in), and may suffer from degradation or inhibition [2].

It is now accepted throughout the world-wide forensic DNA community that a likelihood ratio (LR) approach is required to reliably interpret these types of profiles [3]. Accordingly, recent years have seen a proliferation of probabilistic models, implemented via software, offered to the community as solutions to this problem. Although these probabilistic models rely on different assumptions, and make use of different types of information, they all enable the evaluation of evidence within a LR framework. While these software programs have proven generally useful to facilitate the interpretation of complex DNA profiles, [4–7], no generally accepted guidelines exist to establish their validity for use in forensic casework. Model validation for use in forensic casework is not straightforward because the true weight of the DNA evidence cannot be determined; indeed, the generated LR always depends on the model's assumptions, no 'gold standard' exists in the form of a true likelihood ratio that can serve as a comparison [8,9].

In this paper, we offer a set of definitions and examples that aim to provide guidance in validating software for casework use. We first introduce some general definitions of model and software validation taken from existing fields. We then propose a set of considerations for validating software for forensic use. We illustrate the application with the LRmix program [10], which has been validated for casework use and introduced into a courtroom setting.

2. Definition of validation

Forensic science is not the first discipline to face the challenges of model and software validation. Consequently, it is possible to learn from the experience of scientists working in different fields. We follow Rykiel [11] in his definition of model validation (originally applied to the field of ecological science). This paper is highly cited and is effectively regarded as a 'standard reference'. We regard model validation as a

[*] Corresponding author at: Netherlands Forensic Institute, Netherlands.

http://dx.doi.org/10.1016/j.scijus.2015.11.007
1355-0306/© 2015 The Chartered Society of Forensic Sciences. Published by Elsevier Ireland Ltd. All rights reserved.

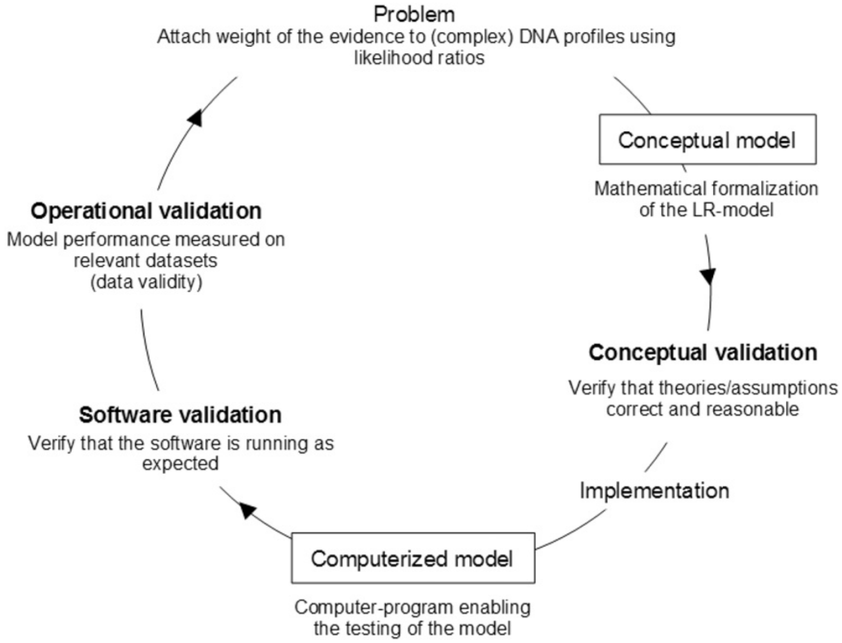

Problem
Attach weight of the evidence to (complex) DNA profiles using likelihood ratios
Conceptual model
Mathematical formalization of the LR-model
Conceptual validation
Verify that theories/assumptions correct and reasonable
Implementation
Computerized model
Computer-program enabling the testing of the model
Software validation
Verify that the software is running as expected
Operational validation
Model performance measured on relevant datasets (data validity)

Fig. 1. Simplified representation of the model development and validation process. The diagram shows the different stages of conceptual, operational and software validation (modified from [11]).

process that results in an explicit statement about the behavior of the model (and subsequently the software). In the case of an interpretation model, such a statement would be: "The implementation of Model X in Software Y is valid for application in forensic casework subject to limitations described in the operational validation document".

Model and software validation are inherently entangled, as software implementation is always needed to implement and use a model (see Fig. 1). However, the two concepts can be related in a simple way; the software is merely a vector for the model. As illustrated in Fig. 1, validated software can actually rely on an invalid model, for example, if the underlying theory or mathematics are shown to be flawed. The goal is to implement a valid model, but it is important to realize that correct implementation of the mathematics of a model by a piece of software provides no information about the validity of the model itself; conversely, demonstration of correct implementation is a critical part of validation.

2.1. Model validation

Model validation ensures that the model has been extensively checked to be sound and fit for purpose. This can be achieved through two steps: conceptual validation and operational validation [11].

2.1.1. Conceptual validation

Conceptual validation verifies that the mathematical formalization of the model, as well as its underlying assumptions, is fundamentally correct. Publication of the theory of the model in peer-reviewed scientific journals allows an opportunity for the underlying theory to be independently assessed, articulates the underlying assumptions, and, most importantly, documents the scientific support for the model structure. For this step to be successful, the model theory must be thoroughly explained. Publication, while necessary, is not sufficient; an editorial decision to publish a paper does not constitute fundamental proof of the scientific validity or usefulness of the contents.

The advent of electronic publication removes space restrictions and allows for the possibility of publishing online supplementary material, and gives modellers the opportunity to expand on their methods. The underlying data on which the conclusions are based can and should be published as supplementary material so that independent researchers can inspect it and use it to independently verify the results obtained. For open-source software, the computer code can also be published as supplementary material, or as a link provided to the location of the code [12]. The code can then be studied by independent researchers, facilitating an understanding of the model, an important component of conceptual validation. The implementation of the model can also then be independently assessed by interested parties.

The most straightforward way to demonstrate conceptual validity is for the model developer to embrace a transparent approach, which allows for true independent review and verification. A transparent approach requires all of the model assumptions to be described, and accessible to anyone who wishes to independently re-implement the model. This approach is demonstrated by [7,8]. This is diametrically opposed to a black-box approach in which only partial explanations are provided, denying an independent researcher the ability to scrutinize the details and re-implement the model if desired [3,13].

2.1.2. Operational validation

We follow [14] and define operational validation as the procedure that determines whether "the model's output behavior has the accuracy required for the model's intended purpose over the domain of the model's intended applicability". Operational validation is usually verified using a "computerized model". In other words, unless a computer implementation of the model is available that can run a profile and yield an output, the operational validity of the model cannot be tested (Fig. 1). Operational validity is tested via user-defined criteria that can be either accepted or rejected. These can be determined for LR-based models. For example, the following properties can readily be tested:

- Comparison to a standard basic model that operates with minimal assumptions so that the effectiveness of models that take into account additional parameters may be measured objectively. Gill and Haned [15] defined the requirements for such model, which allows the evaluation of complex DNA profiles without using all available information.
- The LR of a set of propositions for any profile is lower or equal to the inverse match probability of the profile questioned under the numerator hypothesis [9].
- The LR obtained for a given profile decreases with increasing

ambiguity and decreasing information content [9]. Specifically, any deviation from a one-to-one correspondence of the suspected contributor profile and the evidence profile, as well as any loss of information from the evidence profile itself, should reduce the LR.

- The LR can be compared to a benchmark LR value. A benchmark LR can be calculated when most parameters of the model can be estimated from known profiles (see below). The reasons for any differences between the observed and expected output can be investigated and the model can subsequently be modified to yield the expected output.

2.1.3. Defining benchmarks for LR-based models

Benchmark likelihood ratios can be calculated for certain models for which parameters can be estimated directly from samples with known input. The LRs obtained with parameters estimated from such samples, and the LRs calculated with the estimates for another test dataset should converge [2]. The quality and range of the data used for operational validation are critical [11,14]. We follow Sargent [14] and define data validity as "ensuring that the data necessary for model building, model evaluation and testing, and conducting the model experiments to solve the problem are adequate and correct". Typically, experimental data sets for which the true composition of the samples is known are used (see for example [7,16]). Test samples chosen should represent the spectrum of situations encountered in real-world casework. Profiles representing extreme situations should be included, even if these profiles ultimately might not be interpreted in casework. The idea is to determine not only when the system works as expected, but also when it may fail. Specifically, it is important to investigate the boundaries of the model within its domain of application. Common characteristics of forensic casework samples that can increase their complexity include multiple contributors, low quantity (provoking possible drop-out) and low quality (e.g., degradation, inhibition, contamination). All of these factors increase ambiguity and reduce information content. Both the limitation of the model and the limitations of the evidence must be tested. For example, validation may determine that, past a certain number of contributors, the information content of the profile is simply too limited to reliably distinguish a true contributor from a non-contributor who shares some of the detected alleles by chance. Therefore, based on an operational validation of the model, as implemented by software, it might be relevant to impose a limitation on attempting to interpret casework samples that exceed some defined number of contributors to a mixture. Simulated data can prove helpful in exploring model limitations; however, they cannot substitute for experimental data [13]. Any parameters modeled using simulated data must always be tested on profiles generated from physical samples, and the model refined based on the outcome. The most robust models are those tested with the widest range of data [14]. This is well illustrated by Nordstrom: "The greatest weaknesses of any model computation are the quality of the input data and the adequacy of the assumptions (implicit and explicit); remember GIGO ('garbage in, garbage out')."

2.2. Software validation

Model and software validation usually are carried out simultaneously, as it is the computerized version of the model that enables the model validation exercise (Fig. 1). We define software validation as ensuring that the programmed algorithms follow the mathematical concepts defined in the model. We suggest the following main steps for software validation:

1. Define the statistical specifications of the software: This is an outline of the theory behind the model to be implemented in the computerized version of the model. This document compiles the information that is typically available in peer-reviewed papers describing the model and software implementation.
2. Carry out analytical verification: For example, analytical calculations of likelihood ratios using simple cases (e.g., single-source and two-person mixtures) can be derived and compared to the software output. Depending on the complexity of the model, analytical verification may or may not be possible. This has been termed the "complexity paradox" by [13]; the more complex a model is, the more difficult it is to verify the different blocks of the model. In such a case, the software output can be compared to output from alternative software that implements a similar model.
3. Compare to parallel implementations: Comparisons to alternative software, either relying on a similar or a different probabilistic model, can be useful to verify software behavior. Such comparisons rely on the 'convergence principle' described by Gill and Haned [15], as well as Steele and Balding [1]. Numerical differences between software, corresponding to one unit on the log_{10} scale are negligible [1].
4. Verification of the code itself through visual inspection and recoding. This is most easily achievable through open-source software.

3. Validation in practice

In Box 1 we illustrate an example of validation in practice using the LRmix program, which is freely available in the Forensim R package [17]. While the general approach to validation is applicable to all systems, the specifics will vary depending on the model and the variables that are included. Validation of the model itself will concentrate on the variables that add information content. Questions important to

Box 1
Validation steps for the LRmix program for use in forensic casework.

Step 1. Conceptual validation

- Model theory and assumptions were explained and justified in a "statistical specifications" report,
- Model theory was formalized and published in peer-reviewed journals [18,19].

Step 2. Operational validation

- Model output was compared to expert opinion on 20 cases,
- Model output was compared to the following programs: Lab Retriever [20], LikeLTD [21], FST [16], GRAPE [22],
- Performance tests using 211 controlled mixtures (of one up to five contributors) and 621 (overall-loci) likelihood ratios were compared to expected trends based on gold standard conditions where parameters were known.

Step 3. Software validation

- Software output was evaluated analytically, using the Xcas algebra software
- Model output was evaluated on 77 controlled NGM mixtures, and >1000 LRs were computed and compared to expected trends using known parameters
- LRmix output was evaluated against analytical formulae derived for simple examples
- LRmix output was evaluated against an independent re-implementation of the model (in the Java language), using 77 controlled NGM mixtures, and >1000 LRs were computed and compared

Validation statement: "Over the 1095 LR calculations were submitted to comparisons of LRmix and other software, for all tested samples, the same conclusions were obtained. We therefore concluded that LRmix is validated for use in casework, within the limitations described in the operational validation document."

this process include: how does changing the values for these variables change the LR; at what point does it make a significant difference; what is the effect of using extreme values; and when does the model/software behave in an unexpected way? Box 1 outlines the different steps carried out to validate LRmix for use in forensic casework. These steps are given as an illustration on how the validation procedure might be carried out in casework.

4. Discussion

Although no guidelines yet exist on the best methods to validate forensic DNA interpretation software for casework use, we can draw on the collective wisdom of other disciplines to guide our inquiry. We can also comment on published validation efforts of software tools that have been offered for the interpretation of forensic DNA profiles.

4.1. Does the validation answer the relevant scientific question(s)?

The DNA Commission of the International Society for Forensic Genetics recommendations [3] offers some suggestions for best practice, including a transparent approach that readily allows for independent verification. The Scientific Working Group in DNA Analysis Methods (SWGDAM) has convened a probabilistic genotyping subcommittee that produced a document of concrete guidelines in 2015 [23].

While software validation might appear to be straightforward, model validation may lead to epistemological questions about the true meaning of a validated model. Here, we argue that model validation is possible, given a particular context of use, within a specific framework of limitations, and for an explicit implementation. One of the limitations to consider is the complexity of the model. The more complex the model, the greater the number of assumptions that are required. Increasing the number of variables incorporated into such a model also increases the chance of creating dependencies. Such models require a validation protocol that specifically addresses the additional interactions, and care must be taken to clearly define the variables. We caution that complex models may at some point begin to produce unrealistic results, and hence become counter-productive. More generally, the validation criteria should be explicit to the end users, and a determination made as to whether these criteria are fit for purpose. Within a coherent quality framework, the criteria may be improved over time. As an example, the steps used to validate LRmix are provided to users (Box 1).

4.2. Software and model comparisons

Comparison of the model to be validated to other models is an important part of validation. However, this can be difficult in practice due to differences between the models themselves. Therefore, attempting to set parameters to exactly the same values for each system to perform a fair comparison is not always feasible, as different models rely on different variables and parameters. Typically, imposing values, or including variables, that optimize one system, especially at the expense of other variables important to another system, may produce a misleading comparison.

Such comparisons require careful thought about which variables are important to a model and which, for the sake of the most informative comparison, must be implemented as specified by the model. As a general guideline, external factors, such as the sample population, allele frequencies and, of course, the specific hypotheses compared, can and should be kept constant. However, variables that are used differently between the models, such as analytical threshold, peak heights, dropout model, and even population sub-structure models, must be implemented as originally intended by the architects of the model and software.

4.3. What are the validation responsibilities of the software developer and of the end user?

The extent and type of user validation depends on the credibility of the model and the software implementing it, and this is closely related to the available information (scientific papers, tutorials, websites, seminars and workshops). Any program that uses laboratory-specific data to calibrate input variables requires at least some work on the part of the end-user. Internal validation can also be understood as an important exercise that helps the end-user familiarize himself with the software. An important aspect of this exercise is to identify and understand results that may appear counter-intuitive based on previous experience of the analyst. Assuming the model and software implementation are valid, logical explanations for these results can be found in the details of the calculations. Working through these examples can contribute greatly to the understanding of the scientist [24].

4.4. Is validated software always valid?

Models and software are dynamic; they evolve and improve over time [13]. For example, software validated for STR kits may not be used for SNP markers without an entirely separate validation exercise. This is particularly true for those models relying on empirical data, as such models rely heavily on calibration for their deployment in casework. Casework implementation might also give rise to situations that were not tested during the validation phase; these untested conditions should be submitted to the appropriate validation tests. Similarly, as software are developed or evolve, they can be tested and validated against a repository of simulated and case examples specifically prepared for such a purpose. This ensures that changes to the software are tracked and thoroughly checked.

4.5. The case for transparent software

Adopting a transparent approach is desirable when developing software for use in forensic casework [3,13]. This could be achieved in several ways. An informative discussion on the matter of validation is provided by Nordstrom [13]. Although the author uses examples from geochemistry, we believe the concepts and discussions are also relevant to forensic science. We start with this statement: “Any computer code that is used for regulatory or legal purposes must be transparent” [13]. Freely available, open-source software is a straightforward way to achieve transparency in science, as results obtained with a given software can be verified and reproduced independently [12]. Commercial software can achieve sufficient transparency if the developers choose to provide adequate information about the validation and the performance of the models. It was previously suggested that open-source software can be used as a vehicle to compare the performance of various software, including commercial software [15].

Concerns about the reliability and reproducibility of software used in scientific computing have grown over the last few years [12,25]. There is a strong movement for researchers to make the source code used for analyses freely available to the community at the time of publication. Easily accessible source code implementing a statistical method will allow scientists to perform all aspects of software validation. Availability of code will allow for operational validation as users can apply the method to known samples. Furthermore, independent researchers can visually examine the code to assess the specific implementation of the model. Finally, such transparency will promote standardisation and will facilitate improvements and extensions to existing software which will be a further benefit to the community.

5. Concluding remarks

In 1984, McCarl [24] stated “There is not, and never will be, a totally objective and accepted approach to model validation.” More than

30 years after this statement was made, no generally-accepted method to test the validity of models and algorithms exists, especially in the field of probabilistic genotyping. We hope that the examples and definitions given in this paper will assist both software developers, as well as the end-users in the forensic community, to create and validate interpretation software. It is our hope that the availability of those tools will, in turn, facilitate the introduction LR-based methods for the interpretation of (complex) DNA profiles.

Acknowledgments

PG has received funding support from the European Union Seventh Framework Programme (FP7/2007-2013), Euroforgen-NOE, under grant agreement no. 285487.

References

[1] C.D. Steele, D.J. Balding, Statistical evaluation of forensic DNA profile evidence, Annu. Rev. Stat. Appl. 1 (2014) 361–384.

[2] P. Gill, H. Haned, O. Bleka, B. Hansson, G. Dorum, T. Egeland, Genotyping and interpretation of STR-DNA: low-template, mixtures and database matches—twenty years of research and development, Forensic Sci. Int. Genet. (2015).

[3] P. Gill, L. Gusmão, H. Haned, W. Mayr, N. Morling, W. Parson, L. Prieto, M. Prinz, H. Schneider, P. Schneider, B. Weir, DNA commission of the international society of forensic genetics: recommendations on the evaluation of STR typing results that may include drop-out and/or drop-in using probabilistic methods, Forensic Sci. Int. Genet. 6 (2012) 679–688.

[4] K.E. Lohmueller, N. Rudin, Calculating the weight of evidence in low-template forensic DNA casework, J. Forensic Sci. 58 (2013) S243–S249.

[5] L. Prieto, H. Haned, A. Mosquera, M. Crespillo, M. Alemañ, M. Aler, F. Álvarez, C. Baeza-Richer, A. Dominguez, C. Doutremepuich, M. Farfán, M. Fenger-Grøn, J. García-Ganivet, E. González-Moya, L. Hombreiro, M. Lareu, B. Martínez-Jarreta, S. Merigioli, P.M. del Bosch, N. Morling, M. Muñoz-Nieto, E. Ortega-González, S. Pedrosa, R. Pérez, C. Solís, I. Yurrebaso, P. Gill, Euroforgen-NoE collaborative exercise on LRmix to demonstrate standardization of the interpretation of complex DNA profiles, Forensic Sci. Int. Genet. 9 (2014) 47–54.

[6] C. Benschop, H. Haned, T. de Blaeij, A. Meulenbroek, T. Sijen, Assessment of mock cases involving complex low template DNA mixtures: a descriptive study, Forensic Sci. Int. Genet. 6 (2012) 697–707.

[7] H. Haned, C. Benschop, P. Gill, T. Sijen, Complex DNA mixture analysis in a forensic context: evaluating the probative value using a likelihood ratio model, Forensic Sci. Int. Genet. 16 (2015) 17–25.

[8] D.J. Balding, Evaluation of mixed-source, low-template DNA profiles in forensic science, PNASs 110 (30) (2013) 12241–12246.

[9] R.G. Cowell, T. Graversen, S.L. Lauritzen, J. Mortera, Analysis of forensic DNA mixtures with artefacts, Appl. Stat. 64 (1) (2015) 1–32.

[10] H. Haned, P. Gill, Analysis of complex DNA mixtures using the Forensim package, Forensic Sci. Int. Genet. Suppl. Ser. 3 (2011) e79–e80.

[11] E. Rykiel, Testing ecological models: the meaning of validation, Ecol. Model. 90 (1996) 229–244.

[12] D.C. Ince, L. Hatton, J. Graham-Cumming, The case for open computer programs, Nature 482 (2012) 485–488.

[13] D.K. Nordstrom, Models, validation, and applied geochemistry: issues in science, communication, and philosophy, Appl. Geochem. 27 (2012) 1899–1919.

[14] R.G. Sargent, Verification and validation of simulation models, J. Simulat. 7 (2013) 12–24.

[15] P. Gill, H. Haned, A new methodological framework to interpret complex DNA profiles using likelihood ratios, Forensic Sci. Int. Genet. 7 (2013) 251–263.

[16] A.A. Mitchell, J. Tamariz, K. O'Connell, N. Ducasse, Z. Budimlija, M. Prinz, T. Caragine, Validation of a DNA mixture statistics tool incorporating allelic drop-out and drop-in, Forensic Sci. Int. Genet. 6 (6) (2012) 749–761.

[17] H. Haned, Forensim: an open-source initiative for the evaluation of statistical methods in forensic genetics, Forensic Sci. Int. Genet. 5 (2011) 265–268.

[18] H. Haned, K. Slooten, P. Gill, Exploratory data analysis for the interpretation of low template DNA mixtures, Forensic Sci. Int. Genet. 6 (2012) 762–774.

[19] J.M. Curran, P. Gill, M.R. Bill, Interpretation of repeat measurement DNA evidence allowing for multiple contributors and population substructure, Forensic Sci. Int. 148 (2005) 47–53.

[20] K. Inman, N. Rudin, K. Cheng, C. Robinson, L. Kirschner, A. Inman-Semerau, K. Lohmueller, Lab Retriever: a software tool for calculating likelihood ratios incorporating a probability of drop-out for forensic DNA profiles, BMC bioinformatics 16 (2015) 298.

[21] D. Balding, likeLTD: Likelihoods for Low-Template DNA Profiles, 2012.

[22] S. Grishechkin, K. Prokotfjena, Grape v.3.0, http://www.dna-soft.com/2014 (last retrieved on 20-1-).

[23] Scientific working group on DNA analysis methods, Guidelines for the validation of probabilistic genotyping systems, http://media.wix.com/ugd/4344b0_22776006b67c4a32a5ffc04fe3b56515.pdf2015 (retrieved 25/11/2015).

[24] B.A. McCarl, Model validation: an overview with some emphasis on risk models, Rev. Mark. Agr. Econ. 52 (1984) 153–173.

[25] L.N. Joppa, G. McInerny, R. Harper, L. Salido, K. Takeda, K. O'Hara, D. Gavaghan, S. Emmott, Troubling trends in scientific software use, Science 340 (2013) 814–815.

Exhibit 6

An Addendum to the PCAST Report
on Forensic Science in
Criminal Courts

AN ADDENDUM TO THE PCAST REPORT ON FORENSIC SCIENCE IN CRIMINAL COURTS

On September 20, 2016, PCAST released its unanimous report to the President entitled "*Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods*." This new document, approved by PCAST on January 6, 2017, is an addendum to the earlier report developed to address input received from stakeholders in the intervening period.

Background

PCAST's 2016 report addressed the question of when expert testimony based on a forensic feature-comparison method should be deemed admissible in criminal courts.[1] We briefly summarize key aspects of the previous report.

Forensic feature-comparison methods

PCAST chose to focus solely on forensic feature-comparison methods. These methods seek to determine whether a questioned sample is likely to have come from a known source based on shared features in certain types of evidence. Specific methods are defined by such elements as:

(i) the type of evidence examined (e.g., DNA, fingerprints, striations on bullets, bitemarks, footwear impressions, head-hair);

(ii) the complexity of the sample examined (e.g., a DNA sample from a single person vs. a three-person mixture in which a person of interest may have contributed only 1%); and

(iii) whether the conclusion concerns only "class characteristics" or "individual characteristics" (e.g., whether a shoeprint was made by a pair of size 12 Adidas Supernova Classic running shoes vs. whether it was made by a *specific* pair of such running shoes).

The U.S. legal system recognizes that scientific methods can assist the quest for justice, by revealing information and allowing inferences that lie beyond the experience of ordinary observers. But, precisely because the conclusions are potentially so powerful and persuasive, the law requires scientific testimony be based on methods that are scientifically valid and reliable.[2]

Requirement for empirical testing of subjective methods

In its report, PCAST noted that the *only* way to establish the scientific validity and degree of reliability of a *subjective* forensic feature-comparison method—that is, one involving significant human judgment—is to test it *empirically* by seeing how often examiners actually get the right answer. Such an empirical test of a subjective forensic feature-comparison method is referred to as a "black-box test." The point reflects a central tenet underlying all science: *an empirical claim cannot be considered scientifically valid until it has been empirically tested*.

If practitioners of a subjective forensic feature-comparison method claim that, through a procedure involving substantial human judgment, they can determine with reasonable accuracy whether a particular type of evidence came from a particular source (e.g., a specific type of pistol or a specific pistol), the claim cannot be considered scientifically valid and reliable until one has tested it by (i) providing an adequate number of examiners with an adequate number of test problems that resemble those found in forensic practice and (ii) determining whether they get the right answer with acceptable

[1] As noted in the report, PCAST did not address the use of forensic methods in criminal *investigations*, as opposed to in criminal prosecution in courts.

[2] See discussion of the Federal Rules of Evidence in Chapter 3 of PCAST's report.

frequency for the intended application.[3] While scientists may debate the precise design of a study, there is no room for debate about the absolute requirement for empirical testing.

Importantly, the test problems used in the empirical study define the specific bounds within which the validity and reliability of the method has been established (e.g., is a DNA analysis method reliable for identifying a sample that comprises only 1% of a complex mixture?).

Evaluation of empirical testing for various methods

To evaluate the empirical evidence supporting various feature-comparison methods, PCAST invited broad input from the forensic community and conducted its own extensive review. Based on this review, PCAST evaluated seven forensic feature-comparison methods to determine whether there was appropriate empirical evidence that the method met the threshold requirements of "scientific validity" and "reliability" under the Federal Rules of Evidence.

- In two cases (DNA analysis of single-source samples and simple mixtures; latent fingerprint analysis), PCAST found that there was clear empirical evidence.
- In three cases (bitemark analysis; footwear analysis; and microscopic hair comparison), PCAST found *no empirical studies whatsoever* that supported the scientific validity and reliability of the methods.
- In one case (firearms analysis), PCAST found only one empirical study that had been appropriately designed to evaluate the validity and estimate the reliability of the ability of firearms analysts to associate a piece of ammunition with a specific gun. Because scientific conclusions should be shown to be reproducible, we judged that firearms analysis currently falls short of the scientific criteria for scientific validity.
- In the remaining case (DNA analysis of complex mixtures), PCAST found that empirical studies had evaluated validity within a limited range of sample types.

Responses to the PCAST Report

Following the report's release, PCAST received input from stakeholders, expressing a wide range of opinions. Some of the commentators raised the question of whether empirical evidence is truly needed to establish the validity and degree of reliability of a forensic feature-comparison method.

The Federal Bureau of Investigation (FBI), which clearly recognizes the need for empirical evidence and has been a leader in performing empirical studies in latent-print examination, raised a different issue. Specifically, although PCAST had received detailed input on forensic methods from forensic scientists at the FBI Laboratory, the agency suggested that PCAST may have failed to take account of some relevant empirical studies. A statement issued by the Department of Justice (DOJ) on September 20, 2016 (the same day as the report's release) opined that:

> The report does not mention numerous published research studies which seem to meet PCAST's criteria for appropriately designed studies providing support for foundational validity. That omission discredits the PCAST report as a thorough evaluation of scientific validity.

Given its respect for the FBI, PCAST undertook a further review of the scientific literature and invited a variety of stakeholders—including the DOJ—to identify any "published . . . appropriately designed

[3] The size of the study (e.g., number of examiners and problems) affects the strength of conclusions that can be drawn (e.g., the upper bound on the error rate). The acceptable level of error rate depends on context.

studies" that had not been considered by PCAST and that established the validity and reliability of any of the forensic feature-comparison methods that the PCAST report found to lack such support. As noted below, DOJ ultimately concluded that it had no additional studies for PCAST to consider.

PCAST received written responses from 26 parties, including from Federal agencies, forensic-science and law-enforcement organizations, individual forensic-science practitioners, a testing service provider, and others in the US and abroad.[4] Many of the responses are extensive, detailed and thoughtful, and they cover a wide range of topics; they provide valuable contributions for advancing the field. PCAST also held several in-person and telephonic meetings with individuals involved in forensic science and law enforcement. In addition, PCAST reviewed published statements from more than a dozen forensic-science, law-enforcement and other entities.[5] PCAST is deeply grateful to all who took the time and effort to opine on this important topic.

In what follows, we focus on three key issues raised.

Issue: Are empirical studies truly necessary?

While forensic-science organizations agreed with the value of empirical tests of subjective forensic feature-comparison methods (that is, black-box tests), many suggested that the validity and reliability of such a method could be established *without* actually empirically testing the method in an appropriate setting. Notably, however, none of these respondents identified any *alternative* approach that could establish the validity and reliability of a subjective forensic feature-comparison method.

PCAST is grateful to these organizations because their thoughtful replies highlight the fundamental issue facing the forensic sciences: *the role of empirical evidence*. As noted in PCAST's report, forensic scientists rightly point to several elements that provide critical foundations for their disciplines. However, there remains confusion as to whether these elements can suffice to establish the validity and degree of reliability of particular methods.

(i) The forensic-science literature contains many papers describing variation among features. In some cases, the papers argue that patterns are "unique" (e.g., that no two fingerprints, shoes or DNA patterns are identical if one looks carefully enough). Such studies can provide a valuable *starting point* for a discipline, because they suggest that it may be worthwhile to attempt to develop reliable methods to identify the source of a sample based on feature comparison. However, such studies—no matter how extensive—can *never* establish the validity or degree of reliability of any particular method. Only empirical testing can do so.

(ii) Forensic scientists rightly cite examiners' experience and judgment as important elements in their disciplines. PCAST has great respect for the value of examiners' experience and judgment: they are critical factors in ensuring that a scientifically valid and reliable method is practiced correctly. However, experience and judgment alone—no matter how great—can *never* establish the validity or degree of reliability of any particular method. Only empirical testing of the method can do so.[6]

[4] www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensics_2016__additional__responses.pdf.

[5] www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_forensics_2016_public_comments.pdf.

[6] Some respondents, such as the Organization of Scientific Area Committees' Friction Ridge Subcommittee, suggested that forensic science should be considered as analogous to medicine, in which physicians often treat patients on the basis of experience and judgment even in the absence of established empirical evidence. However, the analogy is inapt. Physicians act with a patient's consent for the patient's benefit. There is no legal requirement, analogous to the requirement imposed upon expert testimony in court by the Federal Rules of Evidence, that physician's actions be based on "reliable principles and methods." Physicians may rely on hunches; experts testifying in court about forensic feature-comparison methods may not.

(iii) Forensic scientists cite the role of professional organizations, certification, accreditation, best-practices manuals, and training within their disciplines. PCAST recognizes that such practices play a critical role in any professional discipline. However, the existence of good professional practices alone—no matter how well crafted—can *never* establish the validity or degree of reliability of any particular method. Only empirical testing of the method can do so.

PCAST does not diminish in any way the important roles of prior research and other types of activities within forensic science and practice. Moreover, PCAST expresses great respect for the efforts of forensic practitioners, most of whom are devoted public servants. It is important to emphasize, however, contrary to views expressed by some respondents, that there is no "hierarchy" in which empirical evidence is simply the best way to establish validity and degree of reliability of a subjective feature-comparison method. In science, empirical testing is the only way to establish the validity and degree of reliability of such an empirical method.

Fortunately, empirical testing of empirical methods is feasible. There is no justification for accepting that a method is valid and reliable in the absence of appropriate empirical evidence.

Issue: Importance of other kinds of studies

In its response to PCAST's call for further input, the Organization of Scientific Area Committees' Friction Ridge Subcommittee (OSAC FRS), whose purview includes latent-print analysis, raised a very important issue:

> While the OSAC FRS agrees with the need for black box studies to evaluate the overall validity of a particular method, the OSAC FRS is concerned this view could unintentionally stifle future research agendas aimed at dissecting the components of the black box in order to transition it from a subjective method to an objective method. If the PCAST maintains such an emphasis on black box studies as the *only* means of establishing validity, the forensic science community could be inundated with predominantly black box testing and potentially detract from progress in refining other foundational aspects of the method, such as those previously outlined by the OSAC FRS, in an effort to identify ways to emphasize objective methods over subjective methods (see www.nist.gov/topics/forensic-science/osac-research-development-needs.) Given the existing funding limitations, this will be especially problematic and the OSAC RFS is concerned other foundational research will thus be left incomplete.

PCAST applauds the work of the friction-ridge discipline, which has set an excellent example by undertaking both (i) path-breaking black-box studies to establish the validity and degree of reliability of latent-fingerprint analysis, and (ii) insightful "white-box" studies that shed light on how latent-print analysts carry out their examinations, including forthrightly identifying problems and needs for improvement. PCAST also applauds ongoing efforts to transform latent-print analysis from a subjective method to a fully objective method. In the long run, the development of objective methods is likely to increase the power, efficiency and accuracy of methods—and thus better serve the public.

In the case of subjective methods whose validity and degree of reliability have already been established by appropriate empirical studies (such as latent-print analysis), PCAST agrees that continued investment in black-box studies is likely to be less valuable than investments to develop fully objective methods. Indeed, PCAST's report calls for substantial investment in such efforts.

The situation is different, however, for subjective methods whose validity and degree of reliability has not been established by appropriate empirical studies. If a discipline wishes to offer testimony based on a subjective method, it must first establish the method's validity and degree of reliability—which can only be done through empirical studies. However, as the OSAC FRS rightly notes, a discipline could follow an alternative path by abandoning testimony based on the subjective method and instead developing an objective method. Establishing the validity and degree of reliability of an objective method is often more straightforward. PCAST agrees that, in many cases, the latter path will make more sense.

Issue: Completeness of PCAST's evaluation

Finally, we considered the important question, raised by the DOJ in September, of whether PCAST had failed to consider "numerous published research studies which seem to meet PCAST's criteria for appropriately designed studies providing support for foundational validity."

PCAST re-examined the five methods evaluated in its report for which the validity and degree of reliability had not been fully established. We considered the more than 400 papers cited by the 26 respondents; the vast majority had already been reviewed by PCAST in the course of the previous study. At the suggestion of John Butler of the National Institute of Standards and Technology (NIST), we also consulted INTERPOL's extensive summary of the forensic literature to identify additional potentially relevant papers.[7] Although our inquiry was undertaken in response to the DOJ's concern, DOJ informed PCAST in late December that it had no additional studies for PCAST to consider.

Bitemark analysis

In its report, PCAST stated that it found no empirical studies whatsoever that establish the scientific validity or degree of reliability of bitemark analysis as currently practiced. To the contrary, it found considerable literature pointing to the unreliability of the method. None of the respondents identified any empirical studies that establish the validity or reliability of bitemark analysis. (One respondent noted a paper, which had already been reviewed by PCAST, that studied whether examiners agree when measuring features in dental casts but did not study bitemarks.) One respondent shared a recent paper by a distinguished group of biomedical scientists, forensic scientists, statisticians, pathologists, medical examiners, lawyers, and others, published in November 2016, that is highly critical of bitemark analysis and is consistent with PCAST's analysis.

Footwear analysis

In its report, PCAST considered feature-comparison methods for associating a shoeprint with a specific shoe based on randomly acquired characteristics (as opposed to with a class of shoes based on class characteristics). PCAST found no empirical studies whatsoever that establish the scientific validity or reliability of the method.

The President of the International Association for Identification (IAI), Harold Ruslander, responded to PCAST's request for further input. He kindly organized a very helpful telephonic meeting with IAI member Lesley Hammer. (Hammer has conducted some of the leading research in the field—including a 2013 paper, cited by PCAST, that studied whether footwear examiners reach similar conclusions when they are presented with evidence in which the identifying features have already been identified.)

[7] The INTERPOL summaries list 4232 papers from 2010-2013 and 4891 papers from 2013-2016, sorted by discipline, see www.interpol.int/INTERPOL-expertise/Forensics/Forensic-Symposium.

Hammer confirmed that no empirical studies have been published to date that test the ability of examiners to reach correct conclusions about the source of shoeprints based on randomly acquired characteristics. Encouragingly, however, she noted that the first such empirical study is currently being undertaken at the West Virginia University. When completed and published, this study should provide the first actual empirical evidence concerning the validity of footwear examination. The types of samples and comparisons used in the study will define the bounds within which the method can be considered reliable.

Microscopic hair comparison

In its report, PCAST considered only those studies on microscopic hair comparison cited in a recent DOJ document as establishing the scientific validity and reliability of the method. PCAST found that none of these studies provided any meaningful evidence to establish the validity and degree of reliability of hair comparison as a forensic feature-comparison method. Moreover, a 2002 FBI study, by Houck and Budowle, showed that hair analysis had a stunningly high error rate in practice: Of hair samples that FBI examiners had found in the course of actual casework to be microscopically indistinguishable, 11% were found by subsequent DNA analysis to have come from different individuals.

PCAST received detailed responses from the Organization of Scientific Area Committees' Materials Subcommittee (OSAC MS) and from Sandra Koch, Fellow of the American Board of Criminalistics (Hairs and Fibers). These respondents urged PCAST not to underestimate the rich tradition of microscopic hair analysis. They emphasized that anthropologists have published many papers over the past century noting differences in average characteristics of hair among different ancestry groups, as well as variation among individuals. The studies also note intra-individual differences among hair from different sites on the head and across age.

While PCAST agrees that these empirical studies describing hair differences provide an encouraging starting point, we note that the studies do not address the validity and degree of reliability of hair comparison as a forensic feature-comparison method. What is needed are empirical studies to assess how often examiners incorrectly associate similar but distinct-source hairs (i.e., false-positive rate). Relevant to this issue, OSAC MS states: "Although we readily acknowledge that an error rate for microscopic hair comparison is not currently known, this should not be interpreted to suggest that the discipline is any less scientific." In fact, this is the central issue: the acknowledged lack of any empirical evidence about false-positive rates indeed means that, as a *forensic feature-comparison method*, hair comparison lacks a scientific foundation.

Based on these responses and its own further review of the literature beyond the studies mentioned in the DOJ document, PCAST concludes that there are no empirical studies that establish the scientific validity and estimate the reliability of hair comparison as a forensic feature-comparison method.

Firearms analysis

In its report, PCAST reviewed a substantial set of empirical studies that have been published over the past 15 years and discussed a representative subset in detail. We focused on the ability to associate ammunition not with a class of guns, but with a specific gun within the class.

The firearms discipline clearly recognizes the importance of empirical studies. However, most of these studies used flawed designs. As described in the PCAST report, "set-based" approaches can inflate examiners' performance by allowing them to take advantage of internal dependencies in the data. The most extreme example is the "closed-set design", in which the correct source of each questioned sample is always present; studies using the closed-set design have underestimated the false-positive and inconclusive rates by more than 100-fold. This striking discrepancy seriously undermines the validity of the results and underscores the need to test methods under appropriate conditions. Other set-based designs also involve internal dependencies that provide hints to examiners, although not to the same extent as closed-set designs.

To date, there has been only one appropriately designed black-box study: a 2014 study commissioned by the Defense Forensic Science Center (DFSC) and conducted by the Ames Laboratory, which reported an upper 95% confidence bound on the false-positive rate of 2.2%.[8]

Several respondents wrote to PCAST concerning firearms analysis. None cited additional appropriately designed black-box studies similar to the recent Ames Laboratory study. Stephen Bunch, a pioneer in empirical studies of firearms analysis, provided a thoughtful and detailed response. He agreed that set-based designs are problematic due to internal dependencies, yet suggested that certain set-based studies could still shed light on the method if properly analyzed. He focused on a 2003 study that he had co-authored, which used a set-based design and tested a small number of examiners (n=8) from the FBI Laboratory's Firearms and Toolmarks Unit.[9] Although the underlying data are not readily available, Bunch offered an estimate of the number of truly independent comparisons in the study and concluded that the 95% upper confidence bound on the false-positive rate in his study was 4.3% (vs. 2.2% for the Ames Laboratory black-box study).

The Organization of Scientific Area Committee's Firearms and Toolmarks Subcommittee (OSAC FTS) took the more extreme position that all set-based designs are appropriate and that they reflect actual casework, because examiners often start their examinations by sorting sets of ammunition from a crime-scene. OSAC FTS's argument is unconvincing because (i) it fails to recognize that the results from certain set-based designs are wildly inconsistent with those from appropriately designed black-box studies, and (ii) the key conclusions presented in court do not concern the ability to sort collections of ammunition (as tested by set-based designs) but rather the ability to accurately associate ammunition with a specific gun (as tested by appropriately designed black-box studies).

Courts deciding on the admissibility of firearms analysis should consider the following scientific issues:

(i) There is only a single appropriate black-box study, employing a design that cannot provide hints to examiners. The upper confidence bound on the false-positive rate is equivalent to an error rate of 1 in 46.

(ii) A number of older studies involve the seriously flawed closed-set design, which has dramatically underestimated the error rates. These studies do not provide useful information about the actual reliability of firearms analysis.

(iii) There are several studies involving other kinds of set-based designs. These designs also involve internal dependencies that can provide hints to examiners, although not to the same extent that closed-set designs do. The large Miami-Dade study cited in the PCAST report and the small studies cited by Bunch fall into this category; these two studies have upper confidence bounds corresponding to error rates in the range of 1 in 20.

From a scientific standpoint, scientific validity should require at least two properly designed studies to ensure reproducibility. The issue for judges is whether one properly designed study, together with

[8] PCAST also noted that some studies combine tests of both class characteristics and individual characteristics, but fail to distinguish between the results for these two very different questions.

[9] PCAST did not select the paper for discussion in the report owing to its small size and set-based design, although it lists it.

ancillary evidence from imperfect studies, adequately satisfies the legal criteria for scientific validity. Whatever courts decide, it is essential that information about error rates is properly reported.

DNA analysis of complex mixtures

In its report, PCAST reviewed recent efforts to extend DNA analysis to samples containing complex mixtures. The challenge is that the DNA profiles resulting from such samples contain many alleles (depending on the number of contributors) that vary in height (depending on the ratios of the contributions), often overlap fully or partially (due to their "stutter patterns"), and may sometimes be missing (due to PCR dropout). Early efforts to interpret these profiles involved purely subjective and poorly defined methods, which were not subjected to empirical validation. Efforts then shifted to a quantitative method called combined probability of inclusion (CPI); however, this approach also proved seriously problematic.[10]

Recently, efforts have focused on an approach called probabilistic genotyping (PG), which uses mathematical models (involving a likelihood-ratio approach) and simulations to attempt to infer the likelihood that a given individual's DNA is present in the sample. PCAST found that empirical testing of PG had largely been limited to a narrow range of parameters (number and ratios of contributors). We judged that the available literature supported the validity and reliability of PG for samples with three contributors where the person of interest comprises at least 20% of the sample. Beyond this approximate range (i.e. with a larger number of contributors or where the person if interest makes a lower than 20% contribution to the sample), however, there has been little empirical validation.[11]

A recent controversy has highlighted issues with PG. In a prominent murder case in upstate New York, a judge ruled in late August (a few days before the approval of PCAST's report) that testimony based on PG was inadmissible owing to insufficient validity testing.[12] Two PG software packages (STRMix and TrueAllele), from two competing firms, reached differing[13] conclusions about whether a DNA sample in the case contained a tiny contribution (~1%) from the defendant. Disagreements between the firms have grown following the conclusion of the case.

PCAST convened a meeting with the developers of the two programs (John Buckleton and Mark Perlin), as well as John Butler from NIST, to discuss how best to establish the range in which a PG software program can be considered to be valid and reliable. Buckleton agreed that empirical testing of PG software with different kinds of mixtures was necessary and appropriate, whereas Perlin contended that empirical testing was unnecessary because it was mathematically impossible for the likelihood-ratio approach in his software to incorrectly implicate an individual. PCAST was unpersuaded by the latter argument. While likelihood ratios are a mathematically sound concept, their application requires

[10] Just as the PCAST report was completed, a paper was published that proposed various rules for the use of CPI. See Bieber, F.R., Buckleton, J.S., Budowle, B., Butler, J.M., and M.D. Coble. "Evaluation of forensic DNA mixture evidence: protocol for evaluation, interpretation, and statistical calculations using the combined probability of inclusion." *BMC Genetics*. bmcgenet.biomedcentral.com/articles/10.1186/s12863-016-0429-7. While PCAST agreed that these rules are *necessary*, PCAST did not review whether these rules were sufficient to ensure reliability and took no position on this question.

[11] The few studies that have explored 4- or 5-person mixtures often involve mixtures that are derived from only a few sets of people (in some cases, only one). Because the nature of overlap among alleles is a key issue, it is critical to examine mixtures from various different sets of people. In addition, the studies involve few mixtures in which a sample is present at an extremely low ratio. By expanding these empirical studies, it should be possible to test validity and reliability across a broader range.

[12] See McKinley, J. "Judge Rejects DNA Test in Trial Over Garrett Phillips's Murder." New York Times, August 26, 2016, www.nytimes.com/2016/08/27/nyregion/judge-rejects-dna-test-in-trial-over-garrett-phillipss-murder.html. The defendant was subsequently acquitted.

[13] Document updated on January 17, 2017.

making a set of assumptions about DNA profiles that require empirical testing.[14] Errors in the assumptions can lead to errors in the results. To establish validity with a range of parameters, it is thus important to undertake empirical testing with a *variety* of samples in the relevant range.[15]

PCAST received thoughtful input from several respondents. Notably, one response[16] suggested that the relevant category for consideration should be expanded from “complex mixtures” (defined based on the number of contributors) to “complex samples” (defined to include also samples with low amounts of template, substantial degradation, or significant PCR inhibition, all of which will also complicate interpretation). We agree that this expansion could be useful.

The path forward is straightforward. The validity of specific PG software should be validated by testing a diverse collection of samples within well-defined ranges. The DNA analysis field contains excellent scientists who are capable of defining, executing, and analyzing such empirical studies.

When considering the admissibility of testimony about complex mixtures (or complex samples), judges should ascertain whether the published validation studies adequately address the nature of the sample being analyzed (e.g., DNA quantity and quality, number of contributors, and mixture proportion for the person of interest).

Conclusion

Forensic science is at a crossroads. There is growing recognition that the law requires that a forensic feature-comparison method be established as scientifically valid and reliable before it may be used in court and that this requirement can only be satisfied by actual empirical testing. Several forensic disciplines, such as latent-print analysis, have clearly demonstrated that actual empirical testing is feasible and can help drive improvement. A generation of forensic scientists appears ready and eager to embrace a new, empirical approach—including black-box studies, white-box studies, and technology development efforts to transform subjective methods into objective methods.

PCAST urges the forensic science community to build on its current forward momentum. PCAST is encouraged that NIST has already developed an approach, subject to availability of budget, for carrying out the functions proposed for that agency in our September report.

In addition, progress would be advanced by the creation of a cross-cutting Forensic Science Study Group—involving leading forensic and non-forensic scientists in equal measure and spanning a range of feature-comparison disciplines—to serve as a scientific forum to discuss, formulate and invite broad input on (i) empirical studies of validity and reliability and (ii) approaches for new technology development, including transforming subjective methods into objective methods. Such a forum would complement existing efforts focused on developing best practices and informing standards and might strengthen connections between forensic disciplines and other areas of science and technology. It might be organized by scientists in cooperation with one or more forensic and non-forensic science organizations—such as DFSC, NIST, IAI, and the American Association for the Advancement of Science.

[14] Butler noted that one must make assumptions, for each locus, about the precise nature of reverse and forward stutter and about the probability of allelic dropout.

[15] Butler noted that it is important to consider samples with different extents of allelic overlap among the contributors.

[16] This response was provided by Keith Inman, Norah Rudin and Kirk Lohmueller.

Exhibit 7

Erin E. Kenneally, Gatekeeping Out of the Box: Open Source Software as a Mechanism to Assess Reliability for Digital Evidence

6 Va. J.L. & Tech 13

Virginia Journal of Law & Technology
Fall 2001

Erin E. Kenneally[a1] [aa1]

GATEKEEPING OUT OF THE BOX: OPEN SOURCE SOFTWARE AS A MECHANISM TO ASSESS RELIABILITY FOR DIGITAL EVIDENCE

“Do you hear that, Mr. Anderson? That is the sound of inevitability.” -Agent Smith, The Matrix

I. Introduction

***1.** It is ironic that scientists studying the voting problems raised in the Florida polls say that the old ways of paper ballots and lever machines give more accurate accounts than punch cards or electronic devices. [1] A study analyzing those problems says that part of the difficulty may lie in voters' lack of familiarity with new technologies. This article addresses a similar issue at the crossroads of law, technology, and science pertaining to standards, reliability, and evidentiary thresholds of proof: digital evidence and the software from which it is generated. Whether or not a vote gets counted or a piece of digital evidence is admitted depends on the standards that are applied to the respective processes. Ultimately it is not the technology itself, but rather the human understanding of its function and capabilities, that paints the picture of truth.

***2.** Just as transparency in the voting scheme and certification steeped in authority go far in resolving allegations of computer glitches, human error, or mischief at the polls, Open Source software may provide a basis for adjudging the reliability of computer experts and programs that promise to be principal sources of evidence in resolving disputes within our technology-dependent society.

***3.** Likewise, just as the focus of poll reform is to establish standards that will prevent the reoccurrence of the recent assault on the process, Open Source software can help anticipate legal challenges that will confront digital evidence, and enable a methodology to assess the credibility of attacks against it.

***4.** Digital evidence has yet to attain widespread smoking gun status. When used to prove claims, it is often a piece or two in the overall puzzle, and questions of its reliability can be quashed by assembling multiple streams of corroborating evidence. Indeed, the implementation of software reliability in industry is dubious at best, and even less clear in the context of the courtroom, where few decisions have squarely addressed more than a cursory evaluation of software reliability. Whether this dearth of judgments can be attributed in part to lack of challenges to the technology that produces digital evidence remains unclear, as there have not been many cases on point to establish precedent. The ubiquity of computer technology that permeates modern society, however, promises to make computer-derived evidence a digital eyewitness. Its reliability must, therefore, be scrutinized in accordance with legal protections.

***5.** This article examines digital evidence reliability by first identifying and differentiating the two competing categories of software from which this evidence is derived: proprietary and Open Source. The next section explores the standards for software reliability in both the industrial marketplace and the legal arena. Specifically, the current standards are addressed in light of their value to industry and the law, as well as their respective historical origins This sets the stage for a reconciliation of standards for reliability as between industry and the courtroom. An outline of the legal approaches to reconciling digital evidence standards and the ensuing dangers of failing to scrutinize the source of the evidence supports the conclusion that the reliability of some digital evidence is not being properly addressed. Finally, this article will advocate the merits of Open Source software as a solution that facilitates the application of appropriate legal standards to novel evidence and helps bridge the gap between the law and industry in measuring reliability.

II. A Software Primer

A. Software Reliability - An Introduction to the Issue

***6.** Why should the reliability of software be a concern to industry and the law collectively? There is a growing tension between the need to present probative and visual evidence of digital disputes and the legal standards for the admissibility of scientific and technical evidence.[2] In other words, the ubiquity of computer technology and pervasiveness of data derived therefrom bear a direct relation to software. Since software provides the functional link between man and machine, unreliable software is sure to have an effect on any activity within the realm of everyday communications, transportation, and even survival within a tightly coupled, high-tech society.[3] The industry that both spawns and is aggrieved by symptoms of unreliable software also drives this issue into the legal arena, where the resolution of many disputes ought to be incumbent upon the reliability of the digital evidence derived from such software. Software reliability, therefore, has immediate and profound meaning in both industry and the courtroom.

***7.** The mention of digital evidence conjures up thoughts of crime scenes, computer cops, and monomaniacal digital bandits - the stuff of "Mission Impossible," handled by the likes of "techies" named Morpheus, and far removed from the concern of Jane Q. Public. In reality, however, digital evidence is not just a byproduct of computer hacker incidents, but is present in virtually any dispute where there exist data relevant to a civil claim or crime. Like other forms of evidence - hair, blood, eyewitness accounts, or paper documents - its form, prevalence, and existence help clarify competing stories about the reenactment of an event. Furthermore, the reliability of both the physical and digital evidence can be scrutinized by examining the forensic science involved in its identification, collection, preservation, and analysis.

***8.** Unlike other forms of evidence where the source is not called into question, however, the reliability of digital evidence hinges on the source being somewhat reliable. That source is the software that generates, processes, and stores data. With blood evidence, for example, we do not question whether Jack Doe's body reliably produces the DNA found in the blood. Either we can DNA-type the blood, or we cannot, and unreliable forensics will not change Jack's DNA into Jill's DNA. With digital evidence, on the other hand, not only will unreliable forensics change the identifying, correlative, and corroborative properties of such evidence, but unreliable software is also capable of maligning the very data upon which a dispute is based or resolved, regardless of exemplary forensic practices.

***9.** Thus, the reliability of software within industry has sobering ramifications for a legal system that requires similar assurances from its evidentiary offers of proof. If industry cannot rely on or has no way of determining whether a piece of software does only what it purports to do, how can courts settle conflicting accounts arising from or backed by this software? One logical conclusion is that the challenge of judging fact from falsehood in litigating disputes will take on the characteristics of an historical fiction novel, albeit one in which discerning truth would be more a function of the attorney's or author's persuasive style than of reference to irrefutable historical records.

***10.** Regardless of whether reliable software has a profit margin, the omnipresent evidence created by software, the relevance of this digital data in virtually every type of legal dispute, and the legal principles governing the admissibility of evidence are subtly affecting industry's need to develop reliability software. How this can be actualized in an industry where software has been and continues to be developed with reliability as an afterthought is where the Open Source model of software development may be useful. This is because the propagation of Open Source software in industry is fueled by the desire for reliable software, which is measured by standards and principles similar to those courts use to determine evidentiary reliability - empirical testing, subjection to peer review and publication, determination of error rate, and general acceptance within the relevant community.

B. Distinguishing Open Source Versus Proprietary Software

***11.** It is important at the outset to identify and differentiate between the two competing categories of software from which digital evidence is derived: proprietary and Open Source. This process is significant because judicial recognition that any “computer-generated” evidence be a product of a “standard” computer program or system in order to gain admission has migrated to issues surrounding software.[4] Furthermore, the distinction is crucial to understanding the nature of the issues involving the reliability of digital evidence, as well as to appreciating the solution offered herein.

***12.** “Standard” computer technology, so far as software is concerned, inexorably refers to COTS (Commercial Off-The-Shelf) products, which are publicly available, ready-made software that can be purchased from a manufacturer.[5] Examples include Microsoft Office, Network Solutions' Check Point RealSecure IDS, Lotus Notes, MacIntosh Operating System, Computer Associates' Inoculate IT, and Norton Anti-virus. The pertinent characteristic of this “standard” software is that it is proprietary, meaning that its underlying source code is not freely available to be viewed or changed.

***13.** In contrast, the features characterizing “nonstandard” software - customizable, able to be manipulated, not necessarily commercially marketed - typify Open Source software. Open access to a computer program's source code is the central defining point for Open Source software, as well as what primarily distinguishes it from proprietary programs. Some prominent examples include RedHat Linux, BIND, and Apache web server software.[6] Open Source is used to describe the conditions under which software source code used by computers is made available to others apart from the developer.[7] Reliability is a significant motivation behind the Open Source movement. This is effectuated by making the code available over the Internet for extensive testing and widespread review to find faults, as well as to catalog responsive code changes, and to maintain concurrent quality control. The software code or some subset can be integrated, packaged, branded, and sold by a developer to customers.[8]

***14.** Proprietary software is often referred to as a “black box” because, without access to the source code, one can only conjecture as to what happens between the data input and data output stages (except, of course, for the programmer and manufacturer). Without access to these “blueprints” a computer professional is left to infer, based on his knowledge and experience, the causes of and solutions to software problems. Open source software demystifies many of these problems by exposing the source code to the public arena. In this way, those with programming skills can “see” what the software is really doing, diagnose problems, and substantiate predicted results of future occurrences.

***15.** The exercise of distinguishing between proprietary and Open Source software for the purposes of evidentiary implications is not intended to paint the former as "bad" and the latter as "good." Nor does the proposed judicial utility of Open Source software preclude economic entitlement, intellectual property protection or technological advancement. [9] Rather, it is intended to stimulate critical analysis of how the legal challenges confronting digital evidence can be overcome by embracing new concepts that comport with evidentiary jurisprudence.

III. Reliability Standards for Software

A. The Importance of Standards to the Law and Industry

***16.** Why is it judicially rational to recognize Open Source software as a mechanism to foster reliability standards for digital evidence? To begin, a standard is any set of conditions that describes a desired or ideal state of something and that can be used to describe or evaluate actual examples of this thing. [10] In the context of software and evidence, standards are one response to the difficulties of evaluating reliability and of mitigating the troubles that accompany imperfect information. [11] Our legal system is guided by notions of reasonableness and judged by objective standards representing society's values. Yet, despite the fact that software is the interface between humans and the machines that permeate modern life, there is no agreement on "specific" processes or standards regarding the reliability of software systems. [12]

***17.** To be sure, there are numerous technical standards being developed for the world wide web, e-commerce and the Internet that have addressed reliability indirectly by focusing on security. [13] History, however, suggests that whenever a major technology or industry has proliferated sufficiently to affect society at large, some measure of social control has followed. [14] This trend has resulted in public health, safety and environmental needs being addressed by laws that incorporate industry-born technical standards by encouraging reliance on them or by adopting those standards by reference. [15] Nevertheless, there is no software Underwriters Laboratory to denote third party certification of reliability; nor is there a software equivalent to the National Electrical Code to provide a common framework from which reliability determinations can be spawned and infused into local codes. Perhaps the closest attempt to duplicate such standards in the software industry has been declared "dead." [16]

***18.** To illustrate this point, one relevant study found 250 different standards applying to the engineering of software, yet observed that these were essentially ineffective, and concluded that software technology was too immature to standardize. [17] Regardless of whether this reasoning holds true at present, the fact remains that software continues to be developed within an alphabet soup of standards. Thus far, many standards that do exist are aimed at promoting interoperability. For example, the well-established "syslog" protocol that allows machines to send event notification messages across Internet Protocol (IP) networks to a central server has become the de facto standard for logging system events. [18] The scalability and flexibility that allow for interoperability between various applications and operating systems, however, are often achieved by sacrificing reliability. [19]

***19.** The legal implications of this discordance of standards in industry are far-reaching. On one level, the utility derived from having standards is comparable between the judicial and industrial arenas. Standards help simplify the decision-making process for determining reliability. In the context of determining the admissibility of evidence, they act as a measuring stick for judges and juries; whereas in industry, they influence the sale of goods for those who are shopping for reliability. Further, standards assuage decision-making for producers of reliability-whether they be developers and vendors in industry, or computer forensics professionals-by narrowing choices. [20]

***20.** Related to this decision-making streamlining is the added benefit of third party validation, which is aimed at providing objectivity and independent assurances. Although the software industry is riddled with authoritative groups like the federal government and independent standards-setting organizations (e.g., IETF, NIST, IISP, ANSI, FIPS[21]), the lack of consensus among this amalgamation has not eased decision-making regarding software reliability. Translated into the courtroom, the law finds no clear authority with which to resolve the issue of “circumstantial guarantees of trustworthiness” of evidence derived from such software.

***21.** Another benefit of accruing standards of reliability for software is grounded in self-perpetuation. The existence of standards invites scrutiny that serves as a check on flaws and accentuates unreliable factors. The underlying rationale is that reliability is enhanced as an end result.[22]

***22.** Most significantly, standards bear directly on the interplay between industry and the legal system. Insofar as the existence of and adherence to software standards can provide protection from legal liability and facilitate claim settlements, the lack of standards can produce the opposite effect. For instance, a legal system bombarded by and ill-equipped to handle issues involving the disputed validity of software-derived evidence (including creation, identification, collection, preservation, and analysis) could result in the admission of evidence the foundation for which is unreliable; the exclusion of digital data that is vital to a claim at hand; the necessity for government regulation to provide clarity; and/or a disincentive to litigate disputes due to tremendous case backlog or, in the best case scenario, escalating pay-to-prove expenses.

1. Industry Standards for Reliable Software

***23.** In order to grasp how the Open Source model can facilitate reliability determinations of digital data within legal standards, it is necessary to understand how industry values and measures software reliability under proprietary models. To begin, software's historical origins, both economic and technical, gave rise to the current disconnect between proprietary software and reliability.

a) History and Origin

***24.** Drawing on the metaphor that organisms are a product of their environment,[23] software[24] reflects the values of its originating environment and the motivations driving its current distribution. Specifically, reliability was not a consideration shaping early software development in the personal computer (PC) environment. This era was characterized by isolated desktops wherein software failures had no way of propagating to other machines, and data had not yet attained its current lifeblood status. Also, errors and malfunctions were an accepted part of the organizational and programming culture, where notions of computer security and intrusions were embryonic. This acceptance was illustrated by wholesale abdication of responsibility by software developers, represented by shrinkwrap licensing.[25]

***25.** In addition, the economic breeding ground for software was not conducive for dissatisfied customers to leverage financial influence upon vendors and developers. This climate resulted, in part, because software could be bought separate from the computer, unlike with its mainframe predecessors. Further, when personal computer (PC) popularity ignited, market share became the key to corporate success and personal financial gain. Market share bore a direct relation to market entry and increased features. Thus, reliability of software became the bastard stepchild to the motivations shaping market share - reduced time spent on testing, and features du jour.[26]

b) Software Reliability Standards at Present

***26.** Against this backdrop of software being developed largely irrespective of reliability, we arrive at the current divergence of standards and dearth of incentives to invest in reliable software. On one hand, lack of accountability in the software industry is a disincentive to producing reliable software. Imagine a society where doctors could use newfangled procedures to treat patients, yet were unencumbered by malpractice liability if someone lost the use of a kidney or reared a child with only one ear. In fact, the computer industry is no stranger to instances of software exposing personal banking records, causing the demise of entire networks, and even facilitating physical harm - oftentimes without matching repercussions upon the developers. [27] Society has become desensitized to the panoply of deficiencies in some software such that customers have become unwitting crash test dummies in the product development cycle. [28]

***27.** In a sense, reliability can be inferred when a party is held responsible if someone/something is not what it claims to be. A doctor's professional vows and integrity, backed by insurance and certifying authorities, provide some objective criteria for gauging reliability in the medical arena. Ultimately, litigation ensures that these measurements are not toothless. This is in stark contrast to the software industry, where developers are motivated by market pressures that traditionally have neither embraced reliability nor penalized unreliability. [29]

***28.** Indeed, the market does not reward reliable software, at least as far as the shortsighted definition of reliability paints it as an impediment to improved functionality, features, and speed. The costs of improving reliability boil down to the production costs of integration and testing. [30] Apparently a piece of software with quick time-to-market, a fancy user interface, and code reminiscent of Swiss cheese is more valuable than one that has undergone significantly more quality control testing to ensure that the hack-of-the-week does not shut down a system.

***29.** This distills another disincentive to invest in reliable software, one rooted in the lack of a widely recognized and measurable definition of reliability. Despite the fact that software customers (read: everyone) are steered by concerns for functionality, consumer choices are at some level based on reliability. For instance, Jack Consumer generally seeks products that are sold widely or by a familiar vendor. In this way there is a propagation of this perception of reliability. This purchasing and production of reliability, however, is occurring within a context of imperfect information. [31] The lack of both agreed upon standards to test reliability and a recognized body to conduct assessments has led to a distortion in the mass consumer market of what reliability means. This deficiency has provided fertile ground for the proliferation of software that addresses well-known, visible, and publicized problems based on maligned information.

***30.** Apart from the social misconception of reliable software, what does software reliability mean to computer industry professionals? Accepting that there are various ways to describe these attributes, reliable software can be characterized by most, if not all, of the following attributes: authenticity (that the data came from or is what it purports to be, and that the software performs as described); integrity (that the data is accurate); availability (that the data is accessible, and that the software does not fail or cause other components to fail); and supportability (that the software has dependable, strong, available support--especially in a production environment). [32] This high-level definition is important in assessing how industry criteria square with legal standards for the reliability of evidence, and ultimately, how the Open Source model may facilitate proper analysis of evidence derived from software.

2. Legal Standards for Reliability

***31.** A basic legal tenet governing the admission of evidence is that it must be relevant, competent, and material to the case at hand. For evidence that is scientific, technical or of a specialized nature, the Federal Rules of Evidence and case law provide standards used by trial courts to determine if such evidence gets admitted.

***32.** The foremost case establishing guidelines for evidentiary reliability was *Daubert v. Merrell Dow Pharmaceuticals* in 1993.[33] *Daubert* involved challenges to the admission of scientific evidence, and aimed to bring clarity to the reliability requirements enunciated in the Federal Rules of Evidence.[34] The guidelines established in *Daubert* require trial courts to consider the following factors in their role as gatekeepers of admissibility:

- Has the scientific theory or technique been empirically tested; or, is it falsifiable?

- Has the theory or technique been subjected to peer review and publication?

- What is the known or potential error rate?

- Is the theory or technique generally accepted within the relevant scientific community?[35]

***33.** These criteria are the Court's attempt to meet the "standard of evidentiary reliability" by ensuring that technical evidence is grounded in knowledge derived from the methods and procedures of science. By tying the validity of the knowledge to the underlying scientific methodology, the Court defines reliability as something that can be validated by testing and supported by more than subjective beliefs or unsupported speculation.[36]

***34.** *Kumho Tire v. Carmichael* extended the *Daubert* guidelines to nonscientific evidence. It gives trial judges the discretion, based on the facts of the case, to determine which *Daubert* criteria should be applied to determine reliability, as well as whether those criteria are satisfied.[37] The Court interpreted the reliability requirement of Federal Rule of Evidence 702 to apply to the word "knowledge," not to "scientific, technical or other specialized."[38] In other words, expert testimony can be based on no scientific knowledge in a particular field in order to meet the standard of evidentiary reliability.

***35.** At first glance, this ruling appears valuable to elucidating standards of reliability for nonscientific evidence. The Court assumes, however, that the four factors devised to measure the reliability of scientific evidence can be applied just as effectively to evaluate technical or specialized evidence. That is to say, the use of scientific methodology to gauge the reliability of scientific knowledge[39] (validation through testing) is the same criterion courts may use to adjudge, for instance, reliability of technical knowledge. If knowledge is based on experience or subjective interpretations that are not susceptible to validation through testing, those factors do not provide assistance in evaluating the reliability of that knowledge.[40]

***36.** Indeed, trial judges are not required to use all of the *Daubert* factors, but may consider them based on the circumstances of a particular case.[41] This flexibility is no less troublesome since it purportedly brings the issue full circle:

trial judges are tasked with gatekeeping the reliability of nonscientific expert testimony, but are left alone to choose which keys will open the lock. This has set the stage for continuing controversy and lack of consistency among jurisdictions in developing and applying standards for admitting nonscientific evidence.[42]

***37.** A judge's handling of novel scientific evidence, for example, is made easier when the underlying knowledge can be readily mapped against the *Daubert* criteria, as is often the case with knowledge rooted in academic or laboratory settings. Whereas, technical knowledge that may accompany the introduction of some forms of digital evidence, generally represents complex and uncharted territory for judges. Coupled with the fact that the claimed expertise may have evolved from an experience-based body of knowledge (i.e. computer forensics, intrusion detection, computer security), the discretion to allow such technical evidence to go before a jury is less predictable.

B. Reconciling Industry and Legal Standards for Reliability

***38.** How does the admissibility of software-derived evidence square with (1) the lack of industry standards for software reliability coupled with (2) unclear legal standards for the reliability of computer-derived evidence?

1. Presumption of Reliability Afforded by Courts?

***39.** It is not an overstatement to say that digital evidence may carry an aura of infallibility in the public's eyes, a fact that may facilitate settlements and discourage technical challenges during litigation. Computer technology is afforded a presumption of reliability because there is a common belief that machines are immune to human frailties, desires, and whims that can lead to erroneous information or misinterpretation.[43] The dangers of placing such unbridled faith in automated systems without appropriate checks, however, are a no less serious matter, as illustrated both within and outside legal contexts.[44]

***40.** Just as the degree of proof necessary to support a chain of custody depends on the nature of the proffered evidence, the mutability of digital evidence makes its digital pathways important in establishing reliability.[45] This approach contrasts with that used to analyze a 19th Century ruby-crested urn that has visibly unique characteristics that are resistant to change, thus making such strict determinations of reliability unnecessary.[46] Similarly, the strictness with which the reliability requirement is applied to software-derived evidence should rest on the importance of the evidence and the extent to which its probative value depends on its accurate and unchanged condition.

***41.** Indeed, the familiar and neat package in which software can display data and the complexities of examining what happens to digital evidence from creation through end product mean that a more watchful eye and steady gavel are needed. The myriad of possibilities contributing to an undetected error in computer-derived evidence includes: errors introduced at one or more of several processing stages; software programmed with errors, programmed to permit errors to go undetected, or programmed to introduce errors into the data; or data rendered inaccurate or biased.[47]

2. Judicial Approaches to Reconciling Reliability of Digital Evidence

***42.** In order to answer the aforementioned question of squaring admissibility with reliability, an assessment of the threshold of scrutiny for computer-derived evidence is needed. The majority of relevant evidentiary challenges have addressed: (a) the authenticity of the evidence; (b) hearsay rule violations and satisfaction of the Business Records Exception; and, (c) the propriety of admitting computer-derived evidence as demonstrative rather than substantive evidence. This is the framework within which judicial discretion has been exercised to control the admission of computer-derived evidence and establish the appropriate standard for evaluating computer-derived evidence.

***43.** Whether or not these thresholds are appropriate for determining the reliability of the underlying software is an important question to which officers of the court should demand a clear answer. Whereas courts are generally aware of the reliability concerns attendant on digital evidence, there have been no reported decisions that squarely address *Daubert* challenges to proprietary commercial software. Reliability determinations have thus far been made indirectly by inquiring, most notably, about the mechanical operation of computer hardware, the accuracy of the human involvement in the data handled by the computer, the commercial availability and use of the program, and the proponent's familiarity with the software. The standards mandated by *Daubert*, *Kumho*, and Federal Rule of Evidence 702, however, demand inquiry into the validity of the scientific theory behind both the specific software technique and the application of those techniques encoded in the software. A more direct and less circumstantial satisfaction of proof calls for analysis of the source code, which is the direct blueprint for a particular piece of software. The proprietary nature of commercial software, however, impedes application of this standard by spurning access to the source code. This has impeded courts from pursuing the easy question, “How reliable is the source code,” and facilitated lower thresholds of scrutiny that seek to prove reliability in circuitous ways.

a) Authenticity Analysis

***44.** Evidence can be admitted by a judge upon the showing of a proper foundation that often entails the establishment of its authenticity and reliability. These preliminary determinations can occur under the auspices of Federal Rule of Evidence 901's requirement that the matter in question is what it is claimed to be, or via the more demanding showing of reliability addressed in F.R.E. 702. The purpose of this initial screening is to ensure that there is a modicum of reliability upon which a jury can then decide what weight that evidence should carry in resolving the issue at hand. The degree of scrutiny applied to determine whether or not computer-derived evidence goes to a jury is unsettled.

***45.** To date, computer-derived data have gained admission upon a foundational showing that the computer process or system produces accurate results when used and operated properly and that it was so employed when the evidence was generated. Federal Rule of Evidence 901 affords a presumption of authenticity to evidence such as x-rays, photographs, tape recordings, computer-generated records or scientific surveys produced by an automated process that is shown to render accurate results. [48] This presumption of reliability has been commonly extended to software performing data storage, collection or retrieval functions. [49] Consequently, a majority of the cases considering the admissibility of such evidence have done so in the context of computerized business records that are maintained or prepared by electronic computing equipment. [50]

***46.** A Texas appellate court, for instance, determined that the computer system that produced and displayed company payroll information, and upon which the expert's testimony was based, was sufficiently authenticated under a state equivalent to F.R.E. 901(b)(9) and did not require an evidentiary hearing. [51] This decision was based on the prosecution's introduction of evidence that the IBM System 38 computer and its programs were in proper working order when they produced the display, that they had done so in the past, and that the technology behind computer-generated displays was trustworthy, reliable, and standard to the computer industry. Further, the systems administrator testified about his years of experience in performing monthly tests on the payroll data system, and stated that the system and its monitoring programs were reliable.

***47.** The irony of this holding underscores the significance of examining the reliability of the software producing the digital evidence. The Defendant was being charged under a violation of the state penal code for harmful access to computer systems. It was alleged that he maliciously modified the source code to his employer's network software to delete a series of computer files bearing critical company information. The defense challenged the reliability of the computer-generated display, which formed the basis of the systems administrator's conclusion that the malicious code caused the deletion of about 160,000 records from payroll. By overruling the objection to the payroll display evidence, while simultaneously upholding Defendant's conviction for manipulating the software that generated such evidence, the Court failed to consider that the software manipulated to produce aberrant results was the same software that was trusted to produce reliable evidence.

***48.** Authentication standards are meant to ensure that the evidence is what it purports to be, and how rigorous a foundation is needed to make this finding depends on the existence of something that can be tested.[52] This is primarily accomplished through the testimony of a witness who can speak to the identity and accuracy of the computer-derived evidence. The rationale is that the availability of a witness who can be cross-examined about the actual event and its link to the digital exhibit is a sufficient guarantee of authenticity.[53] In the context of photographs, for instance, a witness familiar with the picture need only attest that it is a "fair and accurate portrayal" of the scene.[54] Computer-derived evidence has been extended a similar presumption of authenticity by some courts, as long as a computer operator, who is familiar with the process undertaken by the software, testifies.[55]

***49.** The threshold for authenticating computer-derived evidence, however, is ambiguous. Some recommend a higher standard than that applied to photographs,[56] whereas others advocate giving judicial notice of the authenticity of computer-derived evidence under F.R.E. 901(b)(9), which governs authentication of evidence describing a process or system. One of the foremost cases in defining a standard demanded that the proponent of the evidence show the competency of the computer operator; the type of computer used and acceptance in the field as standard and efficient equipment; the procedure for input and output of information, including controls, tests and checks for accuracy and reliability; the mechanical operation of the machine; and the meaning of the records themselves.[57] This rigorous authentication requirement for computer-generated evidence, however, has been eschewed by most courts.[58]

b. The Hearsay Rule and the Business Record Exception

(i) An Analysis

***50.** Another standard to which courts have subjected computer-derived evidence is the evidentiary prohibition against hearsay.[59] It is generally accepted that computer programs that contain out-of-court statements by declarants (computer operators, programmers, data entry personnel) and that are offered to prove the truth of the matter asserted violate the hearsay rule.[60] Nonetheless, federal courts have applied the business records exception (F.R.E. 803(6)) to a wide variety of computer-based information and there is an abundance of case law allowing the introduction of computer-based records under this exception.[61]

***51.** Alternatively, proponents of computer-derived evidence have bypassed hearsay exception hurdles by convincing the court that such evidence constitutes a product of a device performing pre-programmed tasks on admissible data input, as with a radar gun or a calculator.[62] Computerized printouts of phone traces, for example, were not hearsay in one case because they did not rely on the assistance, observations, or reports of a human declarant; the report of phone traces was contemporaneous with the placement of the calls; and the printouts were "merely the tangible result of the computer's internal operations." [63]

***52.** As with authentication and F.R.E. 702 standards, the depth of inquiry and threshold of proof needed to establish computer-derived evidence as a business record are not always clear. One of the earlier cases to address this issue, for example, held computer records inadmissible as business records because of an insufficient foundation. The testimony of a record keeper for the telephone company was insufficient to establish a proper foundation of "trouble recorder cards" at issue because no complete and comprehensive explanation of either their method of preparation or their meaning was provided. This was despite the facts that the witness testified to having direct supervision and control of all the company's records, and that the cards were business records made in the ordinary course of business at or about the times and dates indicated on the cards. [64]

(ii) The Problem with Hearsay and the Business Record Exception as Applied to Digital Evidence

***53.** A basic evidentiary tenet governing admissibility determinations is that there be circumstantial guarantees of trustworthiness to ensure a modicum of reliability so that a jury is not unduly confused or prejudiced by a given piece of evidence. Even though hearsay is generally not allowed because it violates this tenet, factors such as routine reliance on records kept within the normal course of business, lack of motives to fabricate those records, and the non-adversarial nature within which they were created converge to create circumstantial guarantees of trustworthiness. Paper-based business records are occasionally found to be inadmissible if the source of information or the method or circumstances of preparation indicate a lack of trustworthiness. [65]

***54.** Computers may produce different results based on different assumptions of programmers, however, which can only be directly determined by looking at the source code. In other words, these traditional circumstantial guarantees of trustworthiness - that a document or record was made within the normal course of business, at or near the time of the transaction, by someone familiar with the program, etc. - may be irrelevant if the software producing the data is untrustworthy. If the software source code is riddled with bugs [66] vulnerable to exploitation, and/or prone to errors that go undetected by persons relying on that data yet affect the substance of the data deemed to be a business record, the presumption of reliability is faulty. If courts relegate reliability attacks to the coliseum of jurors as an issue affecting the weight of the evidence, is F.R.E. 803(6) being violated, or has the court exposed the factfinder to unfair prejudice or misleading information in violation of F.R.E. 403?

***55.** Even if business records produced by unreliable software are found to satisfy F.R.E. 803(6), one can envision a scenario where F.R.E. 803(7) is used to force the reliability issue. [67] For example, what if Plaintiff accuses Defendant of intruding into his computer system and stealing data. Further, suppose that as part of Plaintiff's computer security policy, Plaintiff maintains and relies on Intrusion Detection System (IDS) software that produces logs from which Plaintiff monitors unauthorized and suspicious activity. Assume that these IDS logs are found to meet the business records exception, yet they do not contain data to support the allegation against Defendant. Can Defendant use the fact that no entries exist in the logs to counter Plaintiff's intrusion claim? In other words, does the absence of a log entry indicate that an intrusion did not occur, or does it merely indicate that the IDS software did not capture the relevant data packets that would have revealed an intrusion?

***56.** Furthermore, in accordance with F.R.E. 803(6), a party against whom the computerized business records are offered must be given the same opportunity to determine its accuracy as is afforded for written business records. [68] Whereas an opponent of a written business record is entitled to inquire about company record-keeping policies and view ledger entries for accuracy, how might that be accomplished when dealing with business records that are produced by software? As with written business records, attempts to alter or delete data can be discerned for the most part, and there is no need to question whether the pen used to mark data entries reliably transferred the information. Computer data manipulations (whether intentional or accidental), however, are easily accomplished without an indication otherwise, and the software creates another dimension between the data input and the digital data compilation sought to be admitted as a business record. This scenario is where examination of the software source code may offer an opponent the chance to determine the record's accuracy if circumstances indicate a lack of trustworthiness.

***57.** Therefore, F.R.E. 803(6)'s ending provision, "... unless the sources of information or other circumstances indicate lack of trustworthiness," although perhaps overshadowed by courts' concern for the other explicit requirements in F.R.E. 803(6), is a reason for divergent rulings on this evidence as well as the evidentiary trump card that harkens a more satisfactory proof of the reliability of computer-derived evidence.

***58.** In addition, it is dangerous to immunize certain computer records from the hearsay rule by likening them to the product of a mechanical process that cannot produce hearsay. It would be persuasive to argue that computer logs, for example, are merely the "tangible result of the computer's internal operations" that do not rely on human observations or reports, and are made contemporaneously with the capturing of data. [69] Unlike phone trace records and calculators, however, the software producing the logs is programmed to capture and process data deemed to be relevant to its programmed function from many computers over a network. Questions about how complete the data capture is and how the logging software decides what should be captured and processed can only be done by examining the underlying source. To admit such evidence without uncovering the assumptions that underlie its function would invite the resolution of claims based on less than a modicum of reliable evidence.

c) Digital Evidence as Demonstrative Analysis

***59.** One final standard used to govern the admission of computer-derived evidence concerns its admission as demonstrative evidence to explain the testimony of a witness. [70] This threshold is wholly deferential to the exercise of judicial discretion, as the standard for review is abuse of discretion. Characterizing evidence as demonstrative can be a strategic attempt to avoid the reliability analysis imposed on substantive evidence so that the evidence can reach a jury. It avoids hearsay attacks and authenticity proofs (because it is not offered for its truth), and dissection under *Daubert* (because it lacks independent probative value and is only offered to illustrate the testimony of an expert). [71]

***60.** It is important to determine whether the software is taking the place of a witness or providing the basis for a witness's testimony, or merely providing a visual portrayal of a witness's verbal testimony that is subject to cross-examination. As with the former situation, admission under the demonstrative evidence standard ignores the need to test assumptions that may exist in the underlying source code and immunizes the software-derived evidence from reliability determinations. In these cases, judges and factfinders encounter an interface dilemma: computer evidence that is presented in a friendly, GUI-fied (graphical user interface) form can hide any number of encoding errors yet have a subliminal impact on issue resolution equivalent to substantive evidence.

***61.** This issue was tangentially addressed in *Perma Research v. Singer*, the seminal case dealing with legal treatment of simulations. [72] Here the Second Circuit overruled the District Court's ruling denying the defense's request to examine a computer program, despite assurances of confidentiality to protect claims of proprietary and "private work product." [73] The defendant argued that the plaintiff's refusal to provide it with the underlying data and theorems of the simulations impaired the defendant's ability to adequately cross-examine the plaintiff's expert witness. [74]

***62.** Could computer logs be deemed a simulation of network activity? An argument could be made that logs are reconstructed images of an event(s) according to input data, which, like any witness testimony, may or may not be credible. Under what standard, for example, should courts admit Microsoft IIS [75] (web server) logs that depict the web activity of any number of users? Is it akin to a chart or graph of a systems administrator's opinion that Jane Doe defaced the former's company's website, and thus subject to a lower threshold? Or are these logs substantive evidence of the alleged incident that should be scrutinized under higher thresholds like *Daubert*? Does it matter that this particular server software was vulnerable to any number of exploits that could have affected the resulting log data? If this is admitted as demonstrative evidence, is its probative value outweighed by its potential to mislead the jury? [76]

***63.** Is testimony about a network intrusion analogous to air crash simulation testimony? Surely courts would be hard pressed to admit airline reconstruction simulation without scrutinizing the quality and quantity of data points contained in the black box. [77] Yet, why should the intrusion detection logs derived from computer network software and presented in a common, browser-like interface be treated any differently?

3. Issues Arising From Attempts to Reconcile Standards

a) The Circumvention of *Daubert*

***64.** As pointed out in the previous section, judicial treatment of computer-derived evidence is the framework within which judges and litigants will most likely manage reliability challenges to the software producing digital evidence. Courts may opt to take judicial notice of the evidence's auhenticity, declare it demonstrative evidence that illustrates the testimony of an expert witness, or admit the digital evidence under the Business Records Exception to the hearsay rule. While these standards have been used to justify the admission of digital evidence in the past, they avoid direct scrutiny of the source of the evidence - the software. Whereas the need to present probative and visual evidence of digital disputes is accelerating, it is occurring under the watch of *Daubert*, *Kumho*, and Federal Rule of Evidence 702. It is clear that *Daubert*, *Kumho*, and Federal Rule of Evidence 702 attach to scientific, technical, or specialized knowledge, so efforts to circumvent software analysis under this standard contravene its purpose. The most obvious solution to facilitate reliability determinations of this type of evidence lies within an already established penumbra of case law surrounding *Daubert*, *Kumho*, and Federal Rule of Evidence 702.

***65.** If computer-derived evidence is admitted under a “fair and accurate portrayal” standard, or is entitled to judicial notice as process or system, are *Daubert* and Federal Rule of Evidence 702 being thwarted?

***66.** As stated above, it is unclear whether the reliability of computer-derived evidence necessitates an evidentiary hearing under *Daubert*. Whether or not this type of evidence is entitled to a presumption of reliability upon such showing or should be scrutinized under the rubric of F.R.E. 702 is unsettled. This issue is beginning to emerge more frequently, undoubtedly fueled by the unclear standards for determining the reliability of technical and specialized knowledge, as well as the case-by-case approach with which F.R.E. 901 has been applied. [78] The case-by-case approach to this issue stems in part from the variety of forms that computer-derived evidence can take (computer-generated business records, computer-stored records, computer logs, animations, re-creations, simulations, etc.). The legal reasoning that has governed admissibility in similar cases is helpful in forecasting how courts may handle future challenges to digital evidence.

***67.** One federal appeals court postulated in dicta that computer evidence resulting from a process or system would not necessitate an evidentiary hearing because the underlying principles of this technology are well understood. Yet "serious questions of accuracy and reliability arise, if at all, only in connection with their application in a particular instance."[79]

***68.** A recent case requiring stricter frontline gatekeeping involved a Washington state court that subjected software used by the state to an evidentiary reliability hearing under *Frye*.[80] In that case, the defense challenged the admissibility of deleted evidence recovered by this forensic software tool. During the suppression hearing, the court asked: (1) is the underlying scientific theory - that deleted files can be recovered - generally accepted within the relative scientific community; and (2) is the technique used - recovery of deleted files - valid?

***69.** The court did not rule on the scientific theory directly, since there was no defense challenge to the existence of deleted data; rather the defendant questioned the data's "completeness." Thus, the challenge really focused on the accuracy of the technique. For the reasons used to deny the defense's challenge to the use of the software, the court inferred the validity of the technique from the commercial availability of the software; the variety of available software so employing this technique; the wide use of those types of software tools/processes by law enforcement and information technology professionals familiar with the targeted, proprietary operating system; and the familiarity and testing of the software tool upon which the expert based his testimony.

***70.** Several implications flow from this decision and its underlying rationale. First, the fact that the court subjected the software tool to a reliability hearing shows judicial recognition of software as a "scientific technique," or, at the very least, as a technique that is suitable and/or prone to the same degree of scrutiny. Further, insofar as this test is utilized to determine the validity of a process used to obtain, enhance, or analyze evidence, it would be difficult to dispute that all software falls under that umbrella. Next, industry usage was decisive in determining admissibility, as the "relevant scientific community" was adjudged to be law enforcement and industry professionals. Finally, the affirmation of findings (deleted files) using other methods and the existence of various programs incorporating this technique (recovery of deleted files) were vital to the software's admissibility. This decision underscores how the judiciary recognizes factors such as repeatability, objectivity, and verifiability in evaluating reliability.

b) Dangers of Presuming Reliability of Proprietary Software

***71.** What lurks behind this low threshold of scrutiny for computer-derived evidence is a dangerous presumption that necessitates critical re-thinking: if the fact that software has become standard and generally available contributes to the de facto reliability of the evidence it generates, is reliability truly being addressed? If not, how would reliability be proven?

***72.** Whatever the level of initial scrutiny, the fact that evidence was spawned by "standard computer industry technology" factored into admissibility determinations. Deference to standard computer industry technology may be a symptom of judges' wide latitude and lack of guidance in handling computer-derived evidence, as well as a foreshadowing of how they will handle challenges to software reliability.

***73.** There are indications that evidence resulting from proprietary software[81] enjoys a presumption of authenticity, while its customizable Open Source counterpart faces a higher hurdle:

> Evidence generated through the use of standard, generally available software is easier to admit than evidence generated with custom software. The reason lies in the fact that the capabilities of commercially marketed software packages are well known and cannot normally be manipulated to produce aberrant results. Custom software, on the other hand, must be carefully analyzed by an expert programmer to insure that the evidence

> being generated by the computer is in reality what it appears to be. Nonstandard or custom software can be made to do a host of things that would be undetectable to anyone except the most highly trained programmer who can break down the program using source codes and verify that the program operates as represented.[82]

***74.** There are several misconceptions here that, if adhered to, may lead to evidentiary standards and precedent constructed on a house of cards. For one, reality is sidestepped by assuming that commercial software is more reliable because its capabilities are well known and cannot be manipulated to produce aberrant results. The distinction between "capabilities" and "intentions" is vital to recognizing that what commercial software is capable of and purports to do are often quite different from what results it produces. Arguably, the industry-wide disclaimers and shrinkwrap licenses are admissions by software developers themselves that reliability is not part of the business plan.

***75.** Another misconceived basis for presuming the authenticity of closed, proprietary software is that it cannot be manipulated. This presumes that the original code harbors no unknown "features"[83] and ensures that deficiencies are not addressed and aberrant results are propagated until and if market pressures force a change. The fact that open software can be made to do a host of things undetectable to none other than a programmer also means that verification that the program fails to operate as represented can be addressed. This is in stark contrast to proprietary software, whose operators proceed on blind faith that the software will perform as advertised despite constant bug fixes, patches, hacks, and vulnerability alerts to correct flaws.[84]

***76.** The application of *Daubert* notwithstanding, at some level the aforementioned standards (see Sec III.B.2, supra) involve the exercise of judicial discretion as to whether such evidence is reliable enough to go before a jury. These standards, however, do not simplify decision-making for judges who must arbitrate conflicting accounts of the reliability of software producing the evidence. This is so because when applied comprehensively in the context of technical disputes involving software-derived evidence, these standards reinforce the call for a mechanism to guide their discretionary judgments of trustworthiness. This has resulted in conflicting rulings on computer-derived evidence that will likely become more ambiguous when challenges to software commence.

***77.** A dangerous precedent is being set if source code for proprietary, commercial software is not required to establish authenticity of computer-processed evidence but such a showing is required for open software. Beside the fact that this distinction is based on faulty presumptions about the nature of proprietary software (as discussed infra), it fails to address the reality that Open Source (custom) software is propagating and becoming a force in the industry. It would be a mistake to assume that several judgments raising the bar for the admission of Open Source-generated evidence will deter the use of custom software. While judge-made law can encourage reliance on technical standards, such standards originate in industry and are influenced by the technical environment. Failure to recognize this course and embrace a mechanism that facilitates reliability determinations regarding evidence that is a guaranteed byproduct of the state of industry will only broaden the disconnect between law and industry.

c) Danger of Inferring Reliability from Market Share

***78.** The preference for standard commercial software carries with it the inference that circumstantial guarantees of trustworthiness are based on market share, which inference has not encouraged the proliferation of reliable software. As discussed *infra*, market share for proprietary software has been attained at the expense of testing and reliability.

***79.** The market share metric infers reliability from acceptance within the user base. Hence, there is a presumption that a particular software program would not be widespread unless it were reliable. The pervasiveness of unreliable software, however, has molded a community of users that has grown increasingly desensitized to unreliable software and its accompanying contraindications as an accepted cost and defining attribute of networked society. This may be attributed to a variety of factors such as lack of accountability demand on vendors, time constraints in the workplace, difficulty collectivizing parties who share the same grievance, and tolerance of vendors' wholesale disclaimers of software performance and quality. Nevertheless, if courts are using widespread commercial deployment to support presumptions of reliability, they are in effect applying market share to measure the "circumstantial guarantees of trustworthiness." Whereas business reliance may have been a proper sieve for dubitable physical documents or data not subject to a panoply of algorithms (i.e., calculators, paper records transferred to computers, simple processing), the realities of software in the modern business world suggest that such deference may be dangerous.

***80.** Furthermore, the traditional business preference for proprietary software is tied back to legal considerations, which are typically as unproven as the software itself. In other words, there is a default preference for proprietary software because it purportedly offers an entity upon which to place blame.[85] Reality proves this to be illusory since commercial vendors are very rarely held accountable when their software does not do what it purports to do. In fact, all commercial proprietary software removes the right to sue the manufacturer.[86] In the absence of objective criteria upon which to hold manufacturers of proprietary software accountable for lack of reliability (i.e., certification, insurance, legal judgments), the presumption that Microsoft's monopoly on the market, for example, denotes reliable software is unfounded.

d) Problems Assessing the Reliability of Proprietary Software

(i) Proprietary Software is Not Amenable to Measuring Reliability Standards

***81.** Perhaps the reason that software reliability standards have been so nebulous is that the development model of proprietary software that dominates the market has been incongruent with measures of reliability. Open Source software provides both a mechanism to facilitate the application of judicial standards, as well as arguably a framework for software development that meets legal thresholds for admissibility. One way to evaluate this predictive solution is by extrapolating the proven features of Open Source software to scientific standards, within the context of legal parameters.

***82.** Verifiability, objectivity and transparency are prominent measurements upon which scientific reliability standards are based. How does one gauge the reliability of closed software? What happens when a machine error obliterates or causes a change in the data being offered as evidence? If there is no trail of crumbs left by human intervention, how does Jack Unsumer dispute machine-generated information produced by software whose code is proprietary and whose effect and outcome cannot be duplicated?

***83.** One option would be to take the developer at its word. Setting veracity aside, how often have courts accepted the subjective testimony of one person, let alone one who has a partial, vested interest in the success of a product? Combine this with the fact that it would be exceptional to find a developer who did not disclaim virtually all guarantees on the functionality of their software, and this option loses any appeal or practicability.

***84.** Another attempt to gauge the reliability of closed software is to base opinions on experience using the software. Experience, however, is only as valuable as the breadth of software use and exposure to disparate circumstances. Large software is often too complex to test in every scenario regardless of the amount of time available. How does this account for silent errors that go undetected because all the possible points of failure have not been tested? Open Source presents less of a risk that the underlying software is unreliable because there is a higher level of testing and feedback. This means more flaws can be discovered during development due to a wider scale input into the architecture and design insight, core program code, and documentation.[87] In addition, the features of commercial, closed software only contribute to system and network complexity, which means that specifications for those components are likely to be incomplete.[88]

***85.** Transparency is another value that characterizes both scientific methodology and the Open Source model. Open Source confers the right to view and modify the workings of a system. Closed software licenses perpetuate any unreliabilities by denying the capacity to make software reliable, such as fixing bugs in code or realizing and correcting integration problems with other packages. Further, proprietary licenses severely impede expectations that those unreliabilities will be alleviated, since there is no guarantee that vendors will fix flaws in a timely manner, if at all.[89] This market force suggestion of which vulnerablities or unreliabilities should be addressed and which ultimately are chosen to be addressed by the developer is reactive. It does not address reliability prior to distribution; the software is still out there and bug fixes are far from uniformly or reliably implemented.[90]

***86.** Finally, quality assurance mechanisms of proprietary software are primarily based on feedback from customers and publicity spin control tactics of the developer. This invites inconsistent levels of quality in different versions within a single COTS product.[91]

(ii) Impracticality of Proving Proprietary Software's Reliability

***87.** Open Source tools and experts employing them help resolve some tensions that accompany efforts to prove the reliability of proprietary software. Postive effects include the cost benefit of encouraging the efficient application of justice, operative circumstantial guarantees of trustworthiness, and protection of litigants' constitutional rights

***88.** The cost involved in proving the reliability of tools or techniques under *Daubert* and expert testimony, however, may create a disincentive to litigation in the civil arena and/or set the stage for a criminal arena that is liable to bankrupt itself.

***89.** Critics argue that the amended F.R.E. 702 will infringe on a litigant's constitutional right to a jury trial by denying parties a fair opportunity to present a complete case or defense, and impose economic barriers to satisfying the standard of proof to the detriment of an economically disadvantaged party.[92]

***90.** This fear may be realized if Open Source software is not embraced in cases where digital evidence is challenged. For instance, if the reliability of logs generated by a commercial software product (Microsoft IIS, for example) is challenged via a motion in limine, a "poor" party may very well have to expend great resources to establish, through documentation and testimony concerning product development, management, testing, and implementation, reliability sufficient to meet this threshold. This process may include procuring employees from the software company to address these questions, and obtaining the proper source code and software development documentation for the particular version at issue. To analogize the latter point, if the reliability of the brake system on a 1990 Jeep Cherokee Sport was at issue, a court would not allow proof of a model year 2000 Jeep Grand Cherokee to suffice.

***91.** The preceding example assumes that the source code and its authors are available and willing to make disclosures. All of those assumptions are unlikely, as it is not uncommon to be unable to locate the original programmers, or for the vendor to lack copies of the source code and accompanying documentation. Furthermore, source code disclosure will not likely be surrendered without first having to overcome claims of privilege. Without the source code, hailing an independent expert is futile, so a party is left with the testimony of an employee of the commercial vendor whose motivations to obscure or offer less-than-enlightening reliability testimony are enormous.

***92.** The bottom line is that a wealthier litigant would have the resources to fund or to raise the bar for such offers of proof. Even if the evidence were to go before a jury, the economic hurdles would not dissipate, but rather, would merely be transumed into the need to convince an entire body of factfinders, rather than a single judge.

***93.** In an alternative scenario where the reliability of Open Source software is challenged, the economic disincentives may be lowered. In this case, the proponent of the evidence is no longer at the mercy of a corporate juggernaut whose source code is kept under lock and key; nor is the proponent relegated to inferring the reliability of the software in question based on personal usage, which may or may not be effective in exposing all of the potential "unreliabilities" present in the software. The problems of proof - and costs associated with same - are able to be diluted in the case of Open Source. Either the proponent will have the technical skills to review the Open Source software himself and make offers of proof; or he will be able to retain any number of objective technicians to look at the code and its attendant details to render an opinion of reliability. In either case, the costs of hunting down an original software developer to testify that his proprietary code does what it purports to do are alleviated.

***94.** Open Source software facilitates scrutiny of the reliability of software. Open Source makes it easier to verify whether or not a vulnerability exists that could have allowed the tampering, alteration, corruption, or forgery of information produced by the software. Although outside the scope of a court's duty, from a proactive angle, there is a greater chance that the Open Source software has been reviewed to discover bugs. Open Source software makes it easier for independent, third parties to verify the existence of bugs and the accuracy of the source code itself because it is maintained by disinterested third parties.

***95.** Another tension that Open Source mollifies is infringement on a litigant's ability to mount a meaningful challenge to admissibility. Normally, courts will disallow challenges to the authenticity of computer-based evidence absent a specific showing that the computer data in question may not be accurate or genuine; mere speculation and unsupported theories generally will not suffice. [93] As with proprietary software, the question arises of what would constitute a specific showing that would cast doubt on the authenticity of computer data. Certainly there are credible data from disinterested third parties documenting the vulnerabilities of specific software. [94] It is questionable whether a court would find that documentation from such sources verifying that the particular version of software at issue was exploitable (e.g., word processing or server software contained a compiled-in back door account with a known password that would allow any local user or remote user able to access a certain port to manipulate any database object on the system) prior to or contemporaneously with the production of the evidence in question. Would fairness dictate that in order to mount a specific showing, the opponent must have the opportunity to examine the source code responsible for the production of the evidence in question?

***96.** With evidence derived from Open Source software, such offers of proof are less problematic since they alleviate the tangential conflict involving proprietary objections, while safeguarding the opponent's ability to substantiate its challenges.

(iii) Changing Technical Environment - Industry and Science Intersect

***97.** Within industry, the technical and economic environment is shifting to recognize the need for reliable software. This shift to embrace reliability and define it in terms similar to scientific precepts illustrates a common ground between industry and science that is elemental to Open Source software theory.

***98.** Software in modern industry is no longer sold or implemented in an isolated, stand-alone PC environment. Instead, computer networks connected by interoperable programs define the digital infrastructure. Data storage forms and techniques have changed. Storage capacities on consumer systems average 20 gigabytes ("GBs"), and data in one document often spans multiple disk surfaces via RAID (Redundant Array of Independent Disks) technology. Peripherals have evolved intelligent modems and network routers. Wide area telecommunication methods are more prevalent, and data and voice convergence is a reality that will become standard in due time.

***99.** Software-created applications have incurred significant changes as well. Software comprising email and client/server applications has not only become ubiquitous, but involves seamless data interfaces overlaying data that are dispersed across a network. Databases are similarly ubiquitous, and the controlling software may assemble a document with one computer from a disbursement among many, which makes its existence quite different than one in a filing cabinet or even on a PC hard drive. Artificial intelligence and computer-directed procedures affect everything, including emergency services, traffic control, manufacturing, sales, and infrastructure management. [95]

***100.** In this environment, reliability can be improved by identifying, anticipating and targeting vulnerabilities and reducing defects. This involves a breadth of understanding about how software interacts with other elements of a larger system. [96] There is a widening gap, however, between the needs of software developers and their ability to evaluate reliability in networked systems. Controlled scientific experimentation under traditional mechanisms (i.e., empirical testing by building the same system repeatedly under controlled conditions, using different approaches) is not feasible due to the cost of building such systems. Therefore, the bulk of laboratory assessments of software reliability are based on testing with small samples that cannot address issues of scaling new technology. [97]

***101.** Furthermore, the pace of information technology development is exceeding the ability of users and software developers to adequately document the workings and content of systems. [98] Microsoft is a prominent example of this black box predicament. The size and complexity of these interactions and continual software changes make software reliability more valuable, yet more challenging to obtain. Ultimately, this difficulty has corresponding evidentiary implications, as the ability to generate, send, receive, store, and process potential digital evidence is likewise evolving. [99]

***102.** As discussed infra (see Sec. III.A.2), judicial scrutiny of scientific or technical evidence is rooted in proving that results produced are repeatable, objective, and verifiable, whereas industry measures the reliability of software in terms of authenticity, integrity, and availability. Open Source software is a mechanism by which both sets of values can be achieved.

***103.** Within industry, Open Source software is purported to be beneficial where (a) reliability, stability, and scalability are highly valued; (b) accurate software design and implementation is not readily verified by means other than peer review; (c) the software is business-critical; or (d) the software facilitates a common computing and communications infrastructure. [100]

***104.** Within legal proceedings, Open Source has a similar effect on the determination of reliability for software-derived evidence. As software increases the discriminatory capabilities of digital evidence, the underlying data must be relied upon as accurate and the software producing it should be relatively immune to failure.

***105.** To be sure, these are probing questions that can be addressed by understanding that what courts are seeking in adjudging evidentiary admissibility is a mechanism by which they can gain circumstantial guarantees of trustworthiness, applied objectively, which mechanism scales to a diversity of cases with minimal potential for error, and that minimizes subjectivity and mitigates error. These standards are why the concept of scientific methodology has been the cornerstone of defining the reliability of evidence. Scientific methodology was founded on principles: the production of objective evidence, the minimization of subjective evidence, and the mitigation of error. Courts have attempted to duplicate these results when dealing with digital evidence, by using the industry standard and/or commercial availability to infer reliability. These measuring sticks,, however, have not been calibrated in the metrics of reliability, but in those of increased features and rapid time-to-market, which work in opposition to reliability.

***106.** Nevertheless, the outlook is not grim in terms of fashioning solutions to deal with complex disputes over the reliability of digital data. The principles and values underlying industry standards are evolving to resemble those of science, as exemplified by industry efforts to make software tools and systems more reliable.[101] Recognizing the convergence of industry's embrace of reliability and the law's requirement of the same, a logical approach would be to embrace the emerging industry standard that is conducive to rendering objective measurements of reliability using scientific principles. The Open Source model of software development embodies the principles of the scientific methodology, and may be the means by which to authenticate or test the reliability of software.[102]

4. Open Source Software As a Solution

***107.** Should evidence obtained through industry standard software be subject to the scientific analysis outlined in *Daubert*? If so, how is this possible with proprietary software? Recognizing that the practical resolution of disputes necessitates that lines be drawn somewhere, imputing the reliability of the systems producing the evidence from widespread use and industry acceptance is not faulty in and of itself.

***108.** Open Source software sits at the crossroads between the need to present probative and visual evidence of digital disputes, and the legal standards beckoned by *Daubert*, *Kumho*, and F.R.E. 702. On one hand, digital evidence that does not square with business record protocols, or is produced by pervasive software whose reliability varies depending on the environment, is pervasive and vital to proving legal claims. On the other hand, legal standards call for specific analysis of scientific, technical or specialized knowledge grounded in assurances of reliability before being presented to a factfinder.

***109.** As precedent indicates, courts have looked to the commercial or standard status of software to draw inferences of reliability. Jurisprudential insistence on measuring reliability via the principles of scientific methodology is skirted when proof of software reliability is based on industry standards that do not embrace those same principles. Where those industry standards are marked as proprietary, closed, and nontransparent, software reliability boils down to successful marketing and distribution, and proof entails a leap of faith that flies in the face of contraindications.[103]

***110.** How can software that, by its own admission, disclaims all but its price tag qualify as an automated process that produces accurate results under F.R.E. 901?

***111.** In a similar vein, courts are wise to scrutinize the computer forensic software techniques called upon after legal mechanisms are invoked to ensure that the “original” data was not altered. Yet, there is an underlying assumption that what is collected accurately reflects the transaction. This reflection becomes a house of mirrors, however, when such judgments are inferred from black boxed coding instructions.

***112.** Courts have been faced with “scientific or technical” evidence generated by computers for some time, such as the printouts from gas chromatographs of chemical substances. Yet analysis has not stopped at determining whether a machine was calibrated and functioning properly before and at the time of testing. Originally, the scientific process underlying drug identification had to be scrutinized using *Daubert*-like factors. Like gas chromatography, for instance, computer software is scientific in that it is based on mathematical algorithms. Yet these same reliability protocols have not been applied consistently and pervasively to "industry standard" software, undoubtedly because it would be impractical under the proprietary regime that defines "industry standards" today. The rapid changes in software technology development mean that the program (or, at least a particular version of the software) would be outdated by the time peer review, publication, and empirical testing could furnish a reliability verdict. In addition, the sheer quantity of testing variables and situations would preclude implementing traditional reliability measurements, because the expense would be prohibitive and serve as a disincentive to litigation.

***113.** Open Source software is one mechanism through which these legal principles can be implemented in modern high-tech society, where industry, science and technology are converging. [104]

a) Open Source: The Digital *Daubert*?

***114.** A comparison of the values and mechanisms defining Open Source with those enumerated by *Daubert* reveals striking commonalities that can be drawn upon to determine the reliability of software-derived evidence.

***115.** Although Open Source is used to describe the conditions under which software source code used by computers is made available to others apart from the developer, its embrace of principles of scientific methodology [105] is what makes it conducive to applying the *Daubert* standard to digital evidence. As pointed out initially, reliability is a significant motivation behind the Open Source movement. The values defining reliability of Open Source are congruent with the values industry has defined, but not yet realized, for reliable software. [106]

(i) Peer Review and Publication

***116.** Similar to the scientific process, peer review plays a central role in the Open Source model. Putting source code out in the open makes it difficult to hide design and implementation flaws that may affect the evidentiary end product. This model allows for independent verification and validation of reliability that is buttressed by the quantity and quality of available functional experts. Unlike with proprietary developers of commercial software, for instance, Open Source contributors do not have the same type of vested interest in the success of the system. Moreover, the software is placed under the scrutiny of hundreds and thousands of independent developers who focus on areas of concern or on areas where they have functional expertise.

***117.** Science and Open Source can both be characterized as adversarial processes. Whereas the former is a search for truth that is played out as competition between ideas using observations and data, the latter is a competition between coders using those same variables to find truth in code and to verify whether the code is accurate (corresponds with what it is intended to do). Peer review is critical to this adversarial process. The Internet and Open Source change control procedures are the hard science journals for software reliability. Opponents might argue that this analogy is unfounded, since software is market-driven, a fact that calls its purity into question. Projects are selected on the basis of there interest to the individuals who participate, however, most of whom do so in their spare time without compensation. The process of selecting Open Source projects is unplanned, and competing implementations and efforts are both welcomed and decentralized. Further, inspiration is not derived from compensation assurances, but from a desire to be part of a movement that embraces the creation of great software or membership in a community with the same values. [107]

***118.** Even if one were to concede the market theory of Open Source software, which holds that all software is driven by economic motives, it should be noted that the adversarial nature of scientific discovery renders it not without fault. Personal gain - whether it be money or fame - is not beyond the purview of scientists. For instance, scientific articles submitted for publication and for funding proposals are not always the best arbiters between competing valid claims because the reviewer is often a competitor for the same resources. Peer review has also been criticized as functioning to perpetuate the current scientific paradigm. [108]

***119.** Finally, peer review is essential to promoting objectivity and minimizing subjectivity by utilizing empirical results to determine and promote reliability. This means that access to code underlying software enables observations and experiments leading to rational conclusions that transcend or minimize individual prejudices and self-interests. [109] A more accurate definition of scientific methodology would be hard to find. Further, this objectivity is only enhanced by the fact that Open Source does not labor under any time-to-market or decisional deadlines that might compromise the objective process.

(ii) Error Rate and Falsifiability

***120.** Open Source provides a mechanism for exposing errors in software. These errors are based on conditions present in the actual use of software. This fact is significant in light of the difficulty of testing software given the increased number and types of systems, features, configurations and implementations; the variety of interactions between different products from different vendors, across different networks; and the divergent interpretations of how software code functions within integrated systems. [110]

***121.** If error rates are only as valid as the corresponding testing procedures, the quality of testing that can be performed on software whose source code is known is significantly higher and more easily verifiable. In other words, the only way to test proprietary, closed source software is by using “black box” techniques, which entail exercising code over a range of inputs and observing the outputs for correctness. [111] This is far from comprehensive given the near-infinite number of possible test cases needed to represent real-world systems and variables. Error rates may be valid for simple check-balancing programs run in environments where all the variables are known, but the same cannot be said for black box testing of far more complex software, such as Intrusion Detection System programs which have significantly more input/output variables. [112]

***122.** Another problem with testing proprietary software using black box techniques is that the correct operation of the software may not be a measurable output. It may, for instance, be easy to verify the correct output of a check-balancing program, which is the account balance, by manually doing the calculations. The same cannot be said, however, for determining if data packets collected over a Gigabit Ethernet were captured by an IDS.

***123.** In addition, black box testing is not comprehensive and is not as “testable.” Without access to the source code, for example, it is impossible to determine if there are portions of the code that have not been executed. If it has not been run, it cannot have been tested. This is likened to having a sleeping bomb in software. [113]

***124.** Lastly, there is empirical evidence that black box testing does not uncover as many errors as a combination of testing methods. [114] The error rate of software under closed source means that this information is kept proprietary, subject to the same claim of privilege that precludes the source code itself from being disclosed. Even so, there is a presumption that error rates of a particular piece of software are known, which would make for an exceptional case since it is well known that testing is the caboose on the time-to-market train. [115]

***125.** Code under the Open Source model is subject to falsification, as modifications in software code to produce desired results can be tested to prove whether the results are reliable. Just as rivals attack the scientific theories at their weakest point, so do other developers/coders scrutinize source code and find flaws in the digital blueprint. Closed source developers are often left to wonder if it was indeed the code in one particular piece of software or some other factor that may have caused a mishap, and they are handicapped by the inability to test and verify their hypotheses by tweaking the code. The best way to actually test a hypothesis is to think through the possible consequences and see if those consequences are observed. If software is coded for X, for example, then one would expect the software to produce Y. If Y results, then the hypothesis - that the code will do what it is coded to do and nothing more - is accurate if or until proven otherwise. [116]

(iii) General Acceptance Within the Relevant Community

***126.** The relevant community in this case would likely refer to those who develop software or have the skill to deal with software at the source code level. Since Open Source is a relatively new concept, measuring its general acceptance would be premature. Those who have an understanding of software programming and the ability to manipulate code to perform desired functions, however, would certainly agree that this feature is preferable to having to contact vendors or draw inferences as to why a program may have crashed, for instance, or how to fix an incompatibility. [117]

***127.** Nonetheless, if utilization by IT industry professionals evidences general acceptance from a practitioner's standpoint, then one need only point out the preference for certain Open Source products over their proprietary counterparts. Open Source Apache web server, for example, is the overwhelming dominant choice of webmasters, as it has held over 50% of the web server market. [118] To exemplify further, Open Source GNU and Linux software were determined to be more reliable than commercial software in a study designed to test failure rates of software utilities. [119]

***128.** Even the software behemoth Microsoft, much maligned for its reputation for monopolizing the software market, has quietly initiated a limited source code sharing program. [120] This endeavor represents a remarkable break from its closed source business model, under which code was only sparingly shared with Micorsoft's most valued customers under strict NDAs and confidentiality contracts. Even though not truly Open Source, insofar as it extends permission to view rather than modify the source code, it demonstrates recognition of the value of having access to the code. By Microsoft's own admission, this access will allow troubleshooters to work backwards to solve software-related problems. Further, the President's Information Technology Advisory Committee (PITAC) has officially recommended that the federal government should encourage the development of Open Source software for development for high end computing. [121]

b. Advantages of Open Source as a Solution

***129.** The arguments for using the Open Source model for determining reliability are compelling given that (1) the issue of software reliability is being forced by the converging meaning and significance of reliability within industry and courts, collectively; and (2) software is oftentimes the singular basis for expert analysis that will give rise to challenges to the digital evidence at all stages of civil or criminal litigation.

***130.** Acknowledging the widespread notion that software is a continuous process rather than a static product, Open Source is responsive to the reality of how applications get used in ways and under circumstances that are not anticipated by the manufacturer.[122] Even the most meticulous coding will not alleviate the potential for software failure and error. It would be unreasonable to expect bulletproof software, as it has been estimated that for every single line of code that is written, X number of bugs is created. Such a reality leads to software that fails to do what it purports to do and/or performs functions unexpectedly.

***131.** The layperson whose Microsoft Excel program crashes, for example, recognizes this truth as "blame-the-computer" grumbling. In such a case, the aggrieved user is left to wonder about the cause of the mishap, feels confounded as to the possibility of future recurrences, and hopes that the work performed has not been irreparably damaged. For the most part, technically skilled professionals have no magic elicir for these problems, since most commercial software conceals its source code - the instructions that interface between the computer user and the computer hardware - under proprietary lock-and-key.

***132.** Essentially, software operates as follows: commands and data are inputted, the software responds under the direction of the underlying code, and output results. Proprietary software is a "black box," because what happens between the input and output stages is anyone's guess (except, of course, the programmer and manufacturer). Without access to these blueprints, a computer professional is left to infer, based on his knowledge and experience, the causes of and solutions to software problems. Open source software demystifies many of these problems by exposing the source code in a public arena. In this way, those with programming skills can "see" what the software is really doing, diagnose problems, and substantiate predictions of future recurrences.

(i) Objectifying the Experts

***133.** Open Source may be valuable to facilitate solutions to issues revolving around experts who will be increasingly hailed to determine the facts of what occurred in complex technical cases, assess critical software application failures or allegations, and render opinions on network intrusions and disruptions.

***134.** The reliability factors enunciated in *Daubert* and F.R.E. 702[123] emanated from a tradition of expert testimony being germane primarily to professionals in academic, laboratory, or scientific settings. Experts testifying to the identification, collection, and analysis of digital evidence do not fit the image of prototypical silver-haired scientists with arms-length credentials. Instead, the breadth of computer science applications and the rapid rate of knowledge creation have expanded the pool of professionals who can assist the factfinder to include the likes of non-degreed twenty-somethings with tongue piercings. Unlike experts in hard sciences like physics and chemistry that are steeped in a rich academic tradition, computer science was not institutionalized within academia until relatively recently. Some of the most technically enlightened professionals can wield no more than a bachelor's degree, yet they possess more knowledge than can be contained in a library of relevant books. Furthermore, the plethora of species of computer science and ever-changing technology make it impossible for any one person to claim expertise in more than a single narrow field. To call someone an expert in computers is analogous to professing someone an expert in medicine.

***135.** Furthermore, much like clinical medicine, computer science exhibits features of both science and art, and expertise is built almost exclusively on experience. Nevertheless, bestowing the designation of "expert" remains largely within the discretion of the judge and partially a function of challenges or the lack thereof from party opponents. Quite often the types of experts who are or will be called to opine on network and diagnostic issues giving rise to challenges to digital evidence do not operate from textbook procedures that have well-known error rates, as is the case with DNA profilers. When trying to determine if there has been a network intrusion as well as when trying to identify the perpetrator, for example, there is no established methodology the deviation from which would help gauge the reliability of an expert systems administrator's testimony. The real challenge lies in being able to discern whether the integration of skills that are developed through experience and practiced as a craft should entitle the witness' testimony to be admitted as that of an expert. A youthful network administrator with a ponytail who works for a dot.com may be as equally qualified to render an opinion as an IT manager who is twice his age and has been employed with General Electric his entire life. By the same token, this awareness creates opportunities for charlatans and those who would have been characterized in the past as "junk scientists," both of whom are motivated by less than noble incentives.

***136.** Indeed, the text of F.R.E. 702 expressly states that an expert may be qualified on the basis of experience, and further recognizes that in certain fields this is the predominant if not sole basis for reliable testimony.[124] *Kumho* maintained that "no one denies that an expert might draw a conclusion from a set of observations based on extensive and specialized experience."[125]

***137.** Open Source software is valuable not so much in helping to determine if a given IT professional is "fit" to be an expert as it is in supporting an infrastructure where the linkage of facts and conclusions about cause and effect opinions of digital evidence can be objectively scrutinized. Thus,unlike expert testimony based on analytical techniques taken from proprietary software, Open Source takes some of the "black art" out of the analysis. Insofar as the Open Source model is more peer-to-peer and widely distributed, it more aptly fits with the disintegration of the hierarchical nature of scientific expert admissibility in the courtroom. No longer is reliability proactively made or reactively judged by professionals at the top of the food chain; rather, the pool of potential independent third party validation is broadened.

(ii) Countering the Automating and "De-skilling" of Experts

***138.** Another reason why Open Source is valuable is because of the emergence and propagation of software that automates the analysis of digital evidence. Such software serves to automate expertise and underscores the importance of having reliable software upon which these analyses rest. Since investigating complex technical cases involves more experience-based testimony, the "facts/data" contemplated in F.R.E. 702 must be scrutinized perhaps more carefully. As discussed, Open Source facilitates the application of *Daubert* to determine the reliability of the software producing the digital data upon which an expert's testimony is based.

***139.** A central question throughout this article concerns the threshold of scrutiny for digital evidence. With respect to expert testimony, one is left to inquire whether the threshold should be raised when it provides the sole basis for an expert opinion about an unwitnessed, digital event. Photographic evidence, for example, is often admitted under the "fair and accurate portrayal" standard. Reality, however, can be altered without varying the image. A photo of an accident scene, for example, can be authenticated by a witness, yet not capture the subtleties such as weather and lighting that would affect the reality of whether or not a sign relevant to an accident could be viewed.

***140.** Likewise, the reality that forms the basis for a technical expert's testimony, such as logs produced by software, can be altered by variances that affect the story that the corresponding data portrays. While a witness can be cross-examined on perception, memory, bias, etc., regarding an accident scene, how can the reliability of a technician's testimony be tested if it is based on results of proprietary black box software? At least with photographs, factfinders can use reason to explore how, given the facts of a case, a scene may have looked at the time of the accident. Conversely, with digital evidence, jurors are completely at the mercy of an expert who may present evidence that they have no way of discerning from the printout created by a monkey at a keyboard.

***141.** Open Source software can facilitate reliability determinations of expert testimony in that it assists in determining if an expert's testimony is based on sufficient facts and data, as well as whether it is the product of reliable principles and methods. This benefit is significant insofar as these digital data, such as computer logs, have become more interpretive and are being used by experts as a basis for their opinions about unwitnessed events, accidents, or crimes.[126] Software tools and systems have become the "digital eyewitnesses" to events that occur over computer networks. The select data points and factual knowledge that they produce, capture and process converge to tell a story about how a digital event may have taken place.

***142.** In light of this automated analysis being performed by software, the line between the human skills needed to apply and those needed to interpret the results produced by the software is quite blurred. In this way, software is fostering the "de-skilling" of experts, as evidenced by integrated suites of software tools programmed to automate the identification, collection, preservation, and analysis of digital evidence. This has engendered "experts" who ask how they can get the software to answer a problem rather than how to find out what is underlying that problem. This is perhaps as dangerous as drawing conclusions about the reliability of software that are not based on knowledge of the source code. Opinions based on Open Source can help gauge the validity of opinions that are drawn from the use of these tools, because the facts and methodologies collectively chosen, employed, and embedded in the software amount to a significant basis for expert testimony.

***143.** It is important to distinguish the use of reliable tools from judging reliable conclusions and/or opinions drawn from the use of those tools and techniques. This distinction explains why the crime scene photographer at a murder scene is not hailed into court nearly as often as the blood spatter expert who uses photographs to draw conclusions about the crime. In the digital crime scene or civil dispute, it is increasingly common for the software not only to take the picture, but to draw conclusions as well. Cross-examining the expert whose opinion is derived from such evidence may call for a cross-examination of the software, which is accurately done by having the source code.

***144.** This fact raises an issue of whether the witness should be qualified for the purposes of establishing a foundation for computer evidence versus opining on technical knowledge under *Daubert* and F.R.E. 702. The former approach treats the testimony under the lower threshold of scrutiny attendant to authentication under F.R.E. 901, whereas the latter would involve a more detailed analysis of the principles and methodologies. The former approach was taken in *People v. Lugashi*, which involved a defense challenge to computer-derived evidence presented in the form of a "data dump."[127] The data dump was a program run on a bank's computer system that retrieved and organized the daily credit card transactions reported to the bank. The defense argued that the bank's systems administrator, who maintained such records and who admitted he was not a computer expert, was incompetent to authenticate the digital data.[128] The court rejected this argument and ruled that "a person who generally understands the system's operation and possesses sufficient knowledge and skill to properly use the system and explain the resultant data, even if unable to perform every task from initial design and programming to final printout, is a 'qualified witness"' for purposes of establishing a foundation for the computer evidence.[129] In so doing, the court impliedly rejected the defense position that only a computer expert who is skilled in programming and could inspect and maintain the software and hardware could testify.[130]

***145.** The issue highlighted by *Lugashi* emphasizes that in order to determine whether testimony should be characterized as expert opinion or lay witness observation depends on whether evidence obtained from a software process is reliable. In order to justifiably afford a presumption of accuracy to the software in question, the sufficiency of the facts and data as well as a determination of whether the software is the product of reliable principles and methods is necessary. In these situations, Open Source software is advantageous for the reasons discussed infra. Only after this foundation has been established should the testimony of a lay witness be allowed to influence admissibility.

(iii) Curbing the Battle of the Experts

***146.** Certain attempts to decide between conflicting accounts of technical issues based on closed source software can be likened to competing expert testimony on the cause of a collapsed bridge based on anything but evidence concerning the girders. In such a case, experts can at most only speculate whether it was destroyed by a bomb, buckled under excess weight, or was ravaged by a horde of mutant beavers.

***147.** Testimony involving Open Source code, however, is insightful. Open Source is valuable in light of Revised F.R.E. 703, which allows experts to opine on inadmissible evidence if it is of a type reasonably relied upon by experts in the particular field. What are the implications if an expert testifies based only on hearsay in logs, and the underlying software producing this log evidence is both proprietary and unreliable? Since various network and system logs are the lifeblood for systems administrators, passing this qualifying portion of F.R.E. 703 would be trivial. An opponent can, however, raise reliability issues of those otherwise inadmissible logs. What such a move does is transplant into the courtroom offers to prove the reliability of the software producing the logs; only this time, the battle of the experts takes center stage in front of the jury. If the source code is in the open, opponents may be able to mount credible and legitimate challenges to its accuracy with more than speculation.[131] In this way, rational decisions that focus on evidentiary weight rather than an expert's appearance and demeanor are more probable.

IV. Conclusion

***148.** In attempting to resolve the issue of reliability of some digital evidence, this article has advocated asking obvious yet probing questions concerning the reliability of the software producing the digital evidence. An outline of the legal approaches to reconciling digital evidence standards and of the ensuing dangers of failing to scrutinize the source of the evidence has supported the conclusion that the reliability of some digital evidence may typically be overlooked. Finally, Open Source software has been offered as a solution for facilitating the application of appropriate legal standards to novel evidence and helping bridge the gap between the law and industry in measuring reliability.

***149.** Evidence derived from proprietary software should not be scrutinized just because the source code is made secret, for closed software is quite often a harbinger of probative, reliable evidence. It is inherently incompatible, however, with the tenets embodied in *Daubert* and F.R.E. 702, and as such creates a dilemma for judges when its reliability is legitimately called into question. Surely, courts can admit digital evidence produced by software of questionable reliability and allow it to go to the weight of the evidence or label it demonstrative evidence. Nevertheless, it is presented before a jury of humans, who cannot delete images from their minds as easily as a computer dumps its memory. As the history of photographic evidence illustrates, the dangers of misleading the jury, creating unfair prejudice, or wasting judicial resources become very real.

***150.** In circumstances where the importance of the evidence and the extent to which its probative value depends on its accuracy demand more than cursory proof, the risk of ceding decision making to automated programs via adherence to institutionalized advocacy mechanisms looms large. In those cases, the predominant question asks what the legal system compromises by continuing to handle issues of digital evidence heedless of *Daubert* standards and the existence of Open Source mechanisms.

Footnotes

[a1] Ms. Kenneally is a Forensic Analyst and Attorney with the San Diego Supercomputer Center, Pacific Institute for Computer Security, where she focuses on substantive research, writing and consultation about prevailing issues related to computer forensics law and computer security in both the criminal and civil arenas. These issues include evidentiary, procedural, and policy implications of new technologies and digital evidence. She has lectured and helped coordinate training conferences for state and local judges, law enforcement personnel and professionals concerned with high technology crime. She holds a Master's of Forensic Sciences degree from George Washington University and a Juris Doctor from Cleveland Marshall School of Law.

[aa1] The author would like to recognize the incalculable guidance and vision provided by Attorney Fred Chris Smith of Santa Fe, New Mexico. Mr. Smith is nationally recognized for his pioneering work in dealing with high technology legal issues. The author is collaborating with Attorney Smith on a series of projects aimed at actualizing the proposals championed in this article and in forthcoming publications.

[1] Reuters News, *Old U.S. Voting Systems May Work Best - Scientists Say*, Findlaw Legal News (February 7, 2001), *available at* http:// www.freerepublic.com/forum/a3a8056331ae5.htm (referencing a study conducted at the California Institute of Technology and the Massachusetts Institute of Technology).

[2] *See* § II.A.2, *infra*.

[3] The computer industry has a long tradition of not being accountable for software directly responsible for killing people. *See* The Risks Digest, Forum on Risks to the Public in Computers and Related Systems ACM Committee on Computers and Public Policy Volume 18: Issue 28 (July 26, 1996), *available at* http://www.infowar.com/iwftp/risks/risks-18/18_28.txt.

[4] *See, e.g.*, Weisman v. Hopf-Himsel, Inc., 535 N.E. 2d 1222, 1226 (Ind. Ct. App. 1st Dist. 1989); People v. Bovioi; Burleson v. State, 802 S.W. 2d at 441; 71 Am. Jur. Trials 111 & 118 (1999).

[5] Although courts have not explicitly used the term “closed source” to refer to “standard” computer programs, there is an implicit assumption that they are interchangeable. At this point in time, the features describing “commercial” and “standard” software are equivalent to “closed source” software. Also, courts have not been presented with issues involving “Open Source” software, and therefore would have no reason to introduce this terminology into their decisions.

[6] BIND is the software that provides the domain name service (DNS) for the entire Internet. For a more comprehensive listing, see http:// www.ceu.fi.udc.es/opensource.org-mirror/docs/products.html.

[7] Eric Raymond, The Cathedral and the Bazaar, (viewed November, 2000) at 20, *available at* http://www.tuxedo.org/<diffesr/writings/cathedral-bazaar/cathedralbazaar-1.html. Although access to a computer program's source code is the major defining point for Open Source software, the distribution terms of what is must be applied in a software license, to be referred to as Open Source, must comply with the following criteria: free redistribution of the program; distribution in source code as well as compiled form; allowance of modifications and derived works; the license must not discriminate against any person or group of persons or field of endeavor; the license must be automatic, no signature required; the rights attached to the program must not depend on the program's being part of a particular software distribution; no restrictions on other software that is distributed along with the licensed software. *See generally*, Bruce Perens, *The Open Source Definition*, Open Sources: Voices From the Open Source Revolution (1999), *available at* http:// www.dibona.com/writing/os/online/perens.html (for a more thorough analysis of the Open Source definition and model).

[8] Raymond, *supra* note 7, at 20.

[9] Certainly, the freedoms and rights attendant on a democratic and capitalistic society justify people's entitlement to protect their intellectual labor and earn its rewards. Moreover, society is driven to develop technology that will automate tasks and “make life more efficient.” For further enlightenment on the economic self-interests and business case for Open Source, *see* Advocacy: The Open Source Case for Business, *at* http:// www.opensource.org/advocacy/case_for_business.html (last visited March 30, 2001).

[10] Academic Press Dictionary of Science and Technology, *available at* http://www.harcourt.com/dictionary/def/9/7/5/7/9757500.html (visited March 14, 2001).

[11] Trust In Cyberspace, Committee on Information Systems Trustworthiness, Computer Science and Telecommunications Board, Commission on Physical Sciences, Mathematics, and Applications, National Research Council, Washington, DC (1998) at 154, *available at* http://www.aci.net/kalliste/tic.htm [hereinafter Trust In Cyberspace].

[12] *Id.* The development of "systematic" processes is the extent of the recognized standard convergence.

[13] *See, e.g.*, http://www.din.de/ni/sc27/doc7.html.

[14] Fallows, *Frontier Days*, Industry Standard, (Nov. 14, 1998), *available at* http:// www.thestandard.com/article/display/0,1151,7618,00.html(visited January 7, 2000).

[15] Email from American Bar Association Section of Science and Technology Law Technical Standardization Law Committee (November, 2000).

[16] *See* Ben Roethke, *The Orange Book is Dead, Long Live the Common Criteria,* Security Wire Digest/Information Security Magazine, (Jan. 18, 2001), *available at* http://www.parallaxresearch.com/news/2001/0122/the_orange_ book.html. *See generally*, Institute of Electrical and Electronics Engineers (IEEE), *Standards*, http://standards.ieee.org/reading/ieee/std/se/12207.0-1996.pdf(visited June 3, 2000). Note that the Department of Defense issued what is referred to as "The Orange Book," which promulgates policy and assigns responsibilities for the security analysis of automatic data processing systems. *See*, *generally,* Department of Defense Trusted Computer System Evaluation Criteria, CSD-STD-001-83, Library No. S225,711 (Aug. 15, 1983), *available at* http://www.cerberussystems.com/INFOSEC/stds/d520028.htm. The DOD intends these criteria for use in the evaluation and selection of ADP systems as it relates to sensitive and classified information. This is regarded as a DOD standard and does not match the non-classified business model for designing and evaluating software and interconnected systems bearing removable media. The time and resource-intensive evaluation process runs counter to the industry-based business models needed to thrive in the competitive markets. Interview with Tom Perrine, Security Manager at the San Diego Supercomputer Center (Aug. 17, 2000).

[17] S.L. Pfleeger, *et al.*, *Evaluating Software Engineering Standards*, IEEE Computer, 27(9): 71-79 (1994).

[18] *See generally*, http://www.ietf.org/Internet-drafts/draft-ietf-syslog-syslog-02.txt.

[19] *Id.* For example, lack of integrity checking of messages such as required time-stamping and delivery verification, lack of authenticity, and lack of confidentiality.

[20] Trust In Cyberspace, *supra* note 11 at 154.

[21] IETF (Internet Engineering Task Force), NIST (National Institute for Standards in Technology) , ANSI (American National Standards Institute), IISP (Information Infrastructure Standards Panel),and FIPS (Federal Information Processing Standards Publications).

[22] Trust In Cyberspace, *supra* note 11 at 199.

[23] Martin Libicki, *The Mesh and The Net: Speculations on Armed Conflict in a Time of Free Silicon*, Institute for National Strategic Studies, http:// www.ndu.edu/ndu/inss/macnair/mcnair28/m028ch06.html (visited November 2000).

[24] The word "software" is used throughout this paper to refer to "proprietary" software unless specifically referred to as "Open Source" software.

[25] Trust In Cyberspace, *supra* note 11 at 72.

[26] *Id.*

[27] *See* Fallows, *supra* note 14; *see generally*, Risks Digest, *available at* http://catless.ncl.ac.uk/Risks.

[28] *See generally*, E. Kenneally, *The Bytes Stop Here: Liability for Negligent Security*, Comp. Security J. (Fall 2000).

[29] Bruce Schneier, Cryptogram (April, 2000), *at* http:// www.counterpane.com/crypto-gram-9904.html.

[30] Trust In Cyberspace, *supra* note 11 at 153 (stating "[E]mphasis on standards can actually be something of an inhibitor. Standards efforts, in their desire to achieve maximum consensus, have very long cycle times (five or more years), which certainly do not fit well with product development and release cycles.").

[31] Trust In Cyberspace, *supra* note at 198.

[32] Interview with Tom Perrine, Security Manager at the San Diego Supercomputer Center (Aug. 17, 2000).

[33] Daubert v. Merrell Dow Phaceuticals, 509 U.S. 572 (1993).

[34] *Id.* (citing Frye v. United States, 293 F. 1013 (D.C. Cir. 1923)) (holding that *Frye* (establishing the standard of general acceptance in the scientific community) was superseded by rule 702 of the Federal Rules of Evidence, which governs expert testimony in federal trials). The issue here concerned the admissibility of scientific evidence (expert testimony based on epidemiological evidence) supporting the claim that the drug Bendectin caused birth defects.

[35] *Id.* at 593-595. Other factors that may be relevant to the consideration include: the relationship of technique to methods established as reliable; the nonjudicial uses of the method; the logical or internal consistency of the hypothesis; the consistency of hypothesis with accepted theories; and, the precision of hypothesis or theory.

[36] *Id.* at 590.

[37] Kumho Tire v. Carmichael, 526 U.S. 137, 147-49 (1999).

[38] *Id.*; *see also Daubert*, 509 U.S. at 589-90.

[39] Observation, statement of proposition to be tested, testing plan devised, selection of methods, testing, interpretation of results to reach hypothesis, generation of predictions based on hypothesis, conclusions based on further testing ... all of which can be re-tested for accuracy or reproduced by other experts in the field to arrive at the same result. *See* John R. Platt, *Strong Interference*, Science, Oct. 16, 1964 at 347-53; *see also* Edward J. Imwinkelried, *The Next Step After* Daubert*: Developing a Similarly Epistemological Approach to Ensuring the Reliability of Nonscientific Expert Testimony*, 15 Cardozo L. Rev. 2271, 2283-84 (1994).

[40] *See generally,* K. Issac deVyver, Comment, *Opening the Door but Keeping the Lights Off:* Kumho Tire Co. v. Carmichael *and the Applicability of the* Daubert *Test to Nonscientific Evidence*, 50 Case W. Res. L. Rev. 177 (1999).

[41] *Kumho*, 526 U.S. at 149-50.

[42] *See* Kristina L. Needham, Note, *Questioning the Admissibility of Nonscientific Testimony After* Daubert*: The Need for Increased Judicial Gatekeeping to Ensure the Reliability of All Expert Testimony*, 25 Fordham Urb. L.J. 541, at 550 (1998); Lynn R. Johnson, et al., *Expert Testimony in Federal Court:* Frye, Daubert, *and* Joiner, A.L.I.-A.B.A. Course of Study Materials, SC33, Feb. 12-13, 1998 (reviewing the application of *Daubert* to nonscientific evidence by appellate courts).

[43] *See generally*, Leon E. White, Article, *Maladjusted Contrivances and Clumsy Automation: A Jurisprudential Investigation*, 9 Harv. J.L. & Tech 375 (1996).

[44] *See*, *e.g.,* Campagna v. Hill, 385 N.Y.S.2d 894, 895 (N.Y. App. Div. 1976). Although this is an antiquated case, the New York State Supreme Court determined that a father was liable for support payments based on computer data that alleged him to be in arrears, despite contrary proof of payment offered in defense. Although the case was reversed, it illustrates the dangers of allowing this aura of infallibility to proceed unabated in judicial decisions.

[45] *See* United States v. Clonts, 966 F.2d 1366 (10th Cir. 1992).

[46] This example is included merely to demonstrate that its obvious, physical uniqueness distinguishes it from digital evidence in terms of susceptibility to alteration from a chain of custody perspective. The issue of authenticity is a separate one, in which case the degree of proof in an art fraud counterfeit claim would more closely resemble the stricter scrutiny threshold that should be considered for many types of digital evidence.

[47] *See generally*, Jerome J. Roberts, *A Practitioner's Primer on Computer-Generated Evidence*, 41 U. Chi. L. Rev. 254, 256 (1974).

[48] 30B Michael H. Graham, Federal Practice and Procedure § 6830 (1975); *see also,* People v. Lugashi, 252 Cal. Rptr. 434 (Cal. Ct. App. 1988) (presuming a data collection software program accurate); People v. Mormon, 422 N.E.2d 1065, 1073 (1981) (presuming a data retrieval program accurate).

[49] *Lugashi*, 205 Cal. Rptr. 454; *Mormon*, 422 N.E.2d at 1073.

[50] *See generally*, Donald Zupanec, Annotation, *Admissibility of Computerized Private Business Records*, 7 A.L.R. 4th (1998) (discussing state and federal cases in which courts have considered whether, and under what circumstances, computerized private business records are admissible as evidence in civil and criminal proceedings). Computer-derived evidence may be admitted via the stipulation of both sides to a case. Oftentimes stipulation occurs when both sides decide to accept or decline to challenge the evidence, whether it comes from ignorance of the technical challenges or disinclination to make a technical defense over a substantive one. Whatever the motivation, this does nothing to enhance the reliability of the evidence or establish precedent to guide other courts and litigants.

Aside from judicial notice of automated processes, the evidentiary challenges to computer-derived evidence have gone to the weight of the evidence, rather than its admissibility. Thus, the evidence may be allowed to go before a jury, but its reliability is contested by attacking the chain-of-custody (human handling) of the digital data. For instance, the computer forensic practices involved in the identification, collection, preservation, and analysis of the digital data are favorite targets for those seeking to discredit evidence.

[51] Burleson v. State, 802 S.W.2d 429 (Tex. Ct. App. 1991).

[52] Fred Galves, *Where the Not So Wild Things Are: Computers in the Courtroom, the Federal Rules of Evidence, and the Need for Institutional Reform and More Judicial Acceptance*, 13 Harv. J. L. & Tech. 161, 230 (2000).

[53] *See id*. at 229.

[54] *See id.*

[55] *See generally*, Donal Zupanec, *supra* note 49.

[56] *See, e.g.,* Edward Hannan, *Computer Generated Evidence: Testing the Envelop*e, 63 Def. Couns. J. 353, 358 (1996)(listing ways to authenticate computer generated evidence).

1. Sources of input data are accurate, reliable, trustworthy in own right (*i.e.*, physical measurements);
2. Assumptions used to quantify non-measured items are reasonable, consistent with laws of nature and bracketed at the upper and lower ends;
3. Commercially recognized hardware is employed;
4. Commercially recognized software used that has capacity of executing applications as intended and subject to appropriate input controls, processing controls, output controls;
5. No relevant data have been overlooked;
6. Data inputted, processed, retrieved by properly trained and supervised technicians.

See also Manual for Complex Litigation (Second) §§ 21.446, at 61-61 (1985) (providing that discovery into the reliability of computerized evidence, "include[ing] inquiry into the accuracy of the underlying source materials, the procedures for storage and processing, and some testing of the reliability of the results obtained," should be conducted "well in advance of trial" The Manual recommends that the proponent must establish, to the court's satisfaction under Uniform Rule of Evidence 104(a), the reliability of the computer equipment used and the data processing techniques applied. This foundation would include expert testimony that the processing programs accurately process the information in the business record database).

[57] Monarch Fed. Sav. & Loan Ass'n v. Genser, 383 A.2d 475 (N.J. Super. Ct. Ch. Div. 1977).

[58] *See, e.g., Lugashi,* 252 Cal. Rptr. 434 (noting that there was no need to require testimony on the reliability of hardware and software of particular computer, or internal maintenance and accuracy tests. But, in that case, there was other, non-digital evidence confirming the claims); *see generally*, Galves, *supra* note 51 at 231.

[59] Various business records exceptions have been codified in various forms, including: Fed R. Evid. 803(6) and A.L.I. Model Code of Evid. 514(1).

The following is not excluded by the hearsay rule:

A memorandum, report, record, or data compilation, in any form, of acts, events, conditions, opinions, or diagnoses, made at or near the time by, or from information transmitted by, a person with knowledge, if kept in the course of a regularly conducted business activity, and if it was the regular practice of that business activity to make the memorandum, report, record or data compilation, all as shown by the testimony of the custodian or other qualified witness ... unless the source of information or the method or circumstances of preparation indicate a lack of trustworthiness.

Fed R. Evid. 803(6) (*emphasis added*).

[60] 22 Charles Alan Wright & Kenneth W. Graham, JR., Federal Practice And Procedure § 5174.1 (Supp. 1998).

[61] "C]omputer data compilations are admissible as business records under Fed. R. Evid. 803(6) if a proper foundation as to the reliability of the records is established." United States v. Briscoe, 896 F.2d 1476, 1494 (7th Cir. 1990) (citing United States v. Croft, 750 F.2d 1354 (7th Cir. 1984)). A survey of federal circuit court decisions reveals that all but three circuits - the Second, Third, and D.C. Circuits - have admitted computer-based documents under Fed. R. Evid. 803(6). *See* United v. Fendley, 522 F.2d 181 186-187 (5th Cir. 1975); United States v. Russo, 480 F.2d 1228, 1240-41 (6th Cir. 1973), *cert. denied*, 414 U.S. 1157 (1974); United States v. Vela, 673 F.2d 86, 89-90 (5th Cir. 1982); United States v. Cestnik, 36 F.3d 904, 909-10 (10th Cir. 1994), *cert. denied*, 513 U.S. 1175 (1995); Ameropan Oil Corp. v. Monarch Air Serv., No. 92 C 3450, 1994 WL 86701, at *3-4, 1994, U.S. Dist. LEXIS 3111 (N.D. Ill. Mar. 14, 1994). *See generally*, Anthony J. Dreyer, Note, *When the Postman Beeps Twice: The Admissibility of Electronic Mail Under The Business Records Exception of the Federal Rules Of Evidence,* 64 Fordham L. Rev. 2285 (1996).

Advisory committee's note to FRE 803(6) points out that "the expression 'data compilation' is used as broadly descriptive of any means of storing information other than conventional words and figures in written or documentary form. It includes, but is by no means limited to, electronic computer storage." The Senate Committee's report on the Federal Rules of Evidence also implicitly recognized that computer-based records fall under the Rule. Fed R. Evid. 803(6).

[62] *See, e.g.*, City of Bellevue v. Lightfoot, 877 P.2d 247 (Wash. Ct. App. 1994).

[63] *See*, *e.g.*, State v. Dunn, 7 S.W.3d 427(Mo. Ct. App. 1999). *See* People v. Holowko, 486 N.E.2d 877, 878-79, 109 Ill. 2d 187, 93 Ill. Dec. 344 (1985); *see also* State v. Armstead, 432 So. 2d 837, 839-41 (La. 1983) (holding that computerized records of phone traces were not hearsay and that such records were computer-generated data to be distinguished from computer-stored declarations).

[64] People v. Gauer, 288 N.E.2d 24, 7 Ill. App. 3d 512 (1972).

[65] *Croft*, 750 F.2d at 1367.

[66] A "bug" is an error in coding or product design that causes computer hardware or software to behave unexpectedly or crash. *See generally* http:// www.bugnet.com.

[67] "Evidence that a matter is not included in the memorandum, reports, records, or data compilations, in any form, kept in accordance with the provisions of paragraph (6), to prove the nonoccurrence or nonexistence of the matter, if the matter was

of a kind of which a memorandum, report, record, or data compilation was regularly made and preserved, unless the sources of information or other circumstances indicate lack of trustworthiness." Fed R. Evid. 803(7).

[68] *See, generally*, *Zupanec*, *supra* note 50 at §§ 3, 7[a]; U.S. v. DeGeorgia, 420 F.2d 889 (1969) (finding that it is immaterial, as far as admissibility is concerned, whether a business record is maintained in a computer rather than company books as long as (1) the opponent is given the same opportunity to inquire into the accuracy of the computer and the input procedures used as he would have to inquire into the accuracy of written business records, and (2) the trial court requires the party offering the computer record to provide a foundation that is sufficient to warrant a finding that the record is trustworthy).

[69] *See Howloko*, 109 Ill. 2d 187, 191.

[70] Mark A. Dombroff, Dombroff on Demonstrative Evidence § 2.27 at 50 (1983) (stating that demonstrative evidence "[is] used to inform the trier of fact of scenes, places, objects and other pertinent data relative to the issues in the litigation that, for numerous reasons, cannot be described with as much force and effect without the use of those aids. The use of diagrams and other visual aids is justified on the grounds that they represent a pictorial reproduction or communication of the senses that may be used in place of descriptive testimony, or simply to supplement such testimony.").

[71] Facts or data relied on by experts need not otherwise be admissible into evidence if the information is "of a type reasonably relied upon by experts in the particular field." Fed R. Evid. 703.

[72] Perma Research and Dev. v. Singer, 542 F.2d 111 (2d Cir. 1976), *cert. denied*, 429 U.S. 987 (1976).

[73] *Id.* (citing Roberts, *A Practitioner's Primer on Computer-Generated Evidence*, 41 U. Chi. L. Rev. 254, 255-256 (1974), ("The possibility of an undetected error in computer-generated evidence is a function of many factors: the underlying data may be hearsay; errors may be introduced in any one of several stages of processing; the computer might be erroneously programmed, programmed to permit an error to go undetected, or programmed to introduce error into the data; and the computer may inaccurately display the data or display it in a biased manner. Because of the complexities of examining the creation of computer-generated evidence and the deceptively neat package in which the computer can display its work product, courts and practitioners must exercise more care with computer-generated evidence than with evidence generated by more traditional means.")).

[74] *Id.* at 115.

[75] Internet Information Server (Microsoft's proprietary webserver software).

[76] The Federal Rules of Evidence provide that relevant evidence should be admitted if its probative value is not outweighed by prejudice, potential to mislead the jury, or excessive consumption of time. *See,* FED. R. EVID. 401-403.

[77] *See also* Haley v. Pan Am. World Airways, 746 F.2d 311 (5th Cir. 1984) (admitting a videotaped simulation of the last moments of Flight 759, before its crash in Louisiana on July 9, 1982, as evidence of pre-impact fear on behalf of the decedent), *reh'g denied*, 751 F.2d 1258 (5th Cir. 1984).

[78] *See generally Zupanec*, *supra* note 50, at §10[a], 11[a].

[79] U.S. v. Downing, 753 F.2d 1224, 1240 n.21 (3rd Cir. 1985).

[80] State of Washington v. Leavell (Okanogan County Cause no. 00-1-0026-8, Telephonic Suppression Hearing, October 20, 2000) (calling for an evidentiary hearing under the *Frye* standard, which was the predecessor to *Daubert* and still remains in effect in a number of states).

[81] Although courts have not explicitly used the term "closed source" to refer to "standard" computer programs, there is an implicit assumption that they are interchangeable. At this point in time, the features describing "commercial" and "standard" software are equivalent to "closed source" software. Also, courts have not been presented with issues involving "Open Source" software, and therefore would have no reason to introduce this terminology into their decisions.

[82] 71 Am. Jur. Trials 111 & 118 (1999).

[83] *See, e.g.*, Ted Bridis, *Microsoft Acknowledges Its Engineers Placed Security Flaw in Some Software*, Wall St. J., April 14, 2000, *available at* http://www.slashdot.org/articles/00/04/14/0619206.shtml (reporting, "Microsoft Corp. acknowledged Thursday that its engineers included in some of its Internet software a secret password--a phrase deriding their rivals at Netscape as 'weenies'--that could be used to gain illicit access to hundreds of thousands of Internet sites world-wide. [...]").

[84] *See generally infra* note 92.

[85] Interview with Thomas Oliver, Security Manager at SAIC (Jan. 15, 2001); series of interviews with Tom Perrine, Manager of Networking and Security at the San Diego Supercomputer (Fall 2000).

[86] *See generally* Kenneally, *supra* note 28; Cem Kaner and David Pels, Bad Software (1998).

[87] *See,* TITLE, *available at* http://www.netaction.org/opensrc/ (date) (offering a qualitative listing of Open Source commentary and leading projects).

[88] Trust In Cyberspace, *supra* note 11 at 72.

[89] *See* Alan Cox, *The Risks of Closed Source Computing*, *available at* http://www.osopinion.com/Opinions/AlanCox/AlanCox1.html; interview with Thomas Oliver, Security Manager at SAIC (Jan. 15, 2001).

[90] Trust In Cyberspace, *supra* note 11 at 72.

[91] *Id.*

[92] *See generally*, *Impact of the New Federal Rules of Discovery on Evidence Practice*, Continuing Education of the Bar California, Business Professional's Network, B-119-154 (January 26, 2001).

[93] *See* U.S. v. Tank, 200 F.3d 627 (9th Cir. 2000); Wisconsin v. Schroeder 613 N.W.2d 911 (2000).

[94] *See generally*, CERT/CC- Computer Emergency Response Team / Coordination Center, *at* http://www.cert.org/; Security Alert for Enterprise Resources, *at* http://www.safermag.com/; Security Focus, *at* http:// www.securityfocus.com; Bugtraq, at http://www.bugtraq.securepoint.com; SANS Security Digest Services, *at* http://www.sans.org/newlook/digests/SAC.htm; Attrition, *at* http://www.attrition.org; Microsoft Technical Updates, Microsoft Security Bulletins, *at* http://www.microsoft.com/technet/security.

[95] *See, e.g.*, Peter Sommer, *Computer Forensics- An Introduction*, http://www.virtualcity.co.uk/vcaforens.htm (last visited March 2, 2001).

[96] Trust In Cyberspace, *supra* note 11 at 67.

[97] *Id.* at 65, 67.

[98] *Id.*

[99] *See* Sommer, *supra* note 94.

[100] *See* Raymond, *supra* note 7 at 17-18.

[101] *See, e.g.*, IEEE Syslog Project, Institute of Electrical and Electronics Engineers (IEEE) (visited June 3, 2000) http:// standards.ieee.org/reading/ieee/std/se/12207.0-1996.pdf; *See*, *generally*, http://www.dacs.dtic.mil/databases/url/key.hts?keycode=2:2450&islowerlevel=1 for a list of education, training and conferences aimed improving software reliability. Most notably is ISACC (International Software Assurance Certification Conference), which aims to "bring together the best

minds in the nation to address the growing concerns with software quality, reliability, and security" http://www.cigital.com/ISACC99/mission.html.

[102] Note, FRE 901(b) consists of illustrations of ways to authenticate evidence. These are only illustrative and not meant to limit the ways authentication can be accomplished. USCS FRE R. 901 (2000) Article IX. Authentication and Identification.

[103] *See* § II.B.4.a.ii (discussing black box testing), *supra.*

[104] *See* § II.B.3.d.*iii, supra*.

[105] Raymond, *supra* note 7 at 20. *See generally* Bruce Perens, The Open Source Definition, in Open Sources: Voices from the Open Source revolution, O'Reilly and Associates, Inc. (1999) http:// eon.law.harvard.edu/property00/alternatives/reading4.htm for a more thorough analysis of the Open Source definition and model).

[106] *See* § II.A.1.b (discussing interoperability, error recovery (bug avoidance and fixing), efficiency, quality, etc.), *supra*.

[107] *See* Jonathan Eunice, Beyond the Cathedral and the Bazaar (May, 1998) http://www.illuminata.com/public/content/cathedral/intro.htm.

[108] See David Goodstein, "*How Science Works*" Reference Manual on Scientific Evidence 2nd Edition, Federal Judicial Center (2000) http:// air.fjc.gov/public/fjcweb.nsf/pages/74 (viewed February 23, 2001).

[109] *Id.*

[110] *See, e.g.*, Bruce Schneier, *Are We Ready for a Cyber-UL?,* ZDNet News Commentary (January 2, 2001) http://www.zdnet.com/zdnn/stories/comment/0,5859,2669708,00.html.

[111] *See* Oliver Cole, *White-Box Testing*, Dr. Dobb's Journal (March 2000) http://www.ddj.com/articles/2000/0003/0003a/0003a.htm; *see generally*, Boris Beizer and Van Nostrand Reinhold, Software TESTING TECHNIQUES, Second Edition, (1990) ISBN 1850328803

[112] *Id.*

[113] *Id.*

[114] *Id. See also* Murray Wood, *et al.*, *Comparing and Combining Software Defect Detection Techniques: A Replicated Empirical Study*, Proceedings of the 6th European Conference Held Jointly with the 5th Acm Sigsoft Symposium on Software Engineering (1997).

[115] *See* § II.B.3.c, *infra*.

[116] Chesterene Cwiklik, *Evaluating Significance and Validity When There Are No Hard Numbers,* National Conference on Science and the Law, San Diego, CA (October 11, 2000) at 5.

[117] The author bases this statement on the cumulative discussions with computer science professionals both at the San Diego Supercomputer Center and within industry.

[118] *See* http://www.ceu.fi.udc.es/opensource.org-mirror/docs/products.html.

[119] Reliability testing here was as measured by occurrence of hangs (loops indefinitely) or crashes with a core dump of Open Source GNU and Linux software as compared to commercial software. Specifically, the failure rate of public GNU utilities was lowest in study at 6% versus 15-43% for utilities on commercial versions of UNIX from 5 vendors. ftp:// grilled.cs.wisc.edu/technical_papers/fuzz-revisited.pdf.

[120] *See* Todd Weiss, *Microsoft to Expand Windows source-code sharing program,* Computerworld (February 2, 2001) http://www.computerworld.com/cwi/stories/0,1199,NAV47-68-84-88-93_STO57316,00.html.

[121] *See Recommendations of the Panel on Open Source Software for High End Computing*, President's Information Technology Advisory Committee (September, 2000). *See*, *generally,* Marcus Maher, *Open Source Software: The Success of an Alternative Intellectual Property Incentive Paradigm*, 10 Fordham I.P., Media & Ent. L.J. 619 (Spring 2000).

[122] *See* § II.A.1.b, *supra*.

[123] Fed. R. Evid 702: If scientific, technical, or other specialized knowledge will assist the trier of fact to understand the evidence or to determine a fact in issue, a witness qualified as an expert by knowledge, skill, experience, training, or education, may testify thereto in the form of an opinion or otherwise, if (1) the testimony is based upon sufficient facts or data, (2) the testimony is the product of reliable principles and methods, and (3) the witness has applied the principles and methods reliably to the facts of the case.

[124] *See, e.g.,* U.S. v. Jones, 107 F.3d 1147 (6th Cir. 1997) (admitting handwriting examiner testimony based on years of practical experience, training, and his detailed explanation of his methodology); Tassin v. Sears Roebuck, 946 F.Supp. 1241, 1248 (M.D.La. 1996) (admit ting testimony of design engineer when opinion based on facts, reasonable investigation, traditional technical/mechanical expertise, and a reasonable link between the information and conclusions).

[125] *Kumho,* 526 U.S. at 1178.

[126] Samuel Guiberson, *Panel IV. Scientific and Demonstrative Evidence, Is Seeing Believing*? Proceedings-National Conference of Science and the Law at 163 (July 2000).

[127] *Lugashi*, 205 Cal. App. 3d 632.

[128] *Id.* at 640.

[129] *Id.*

[130] *Id.*

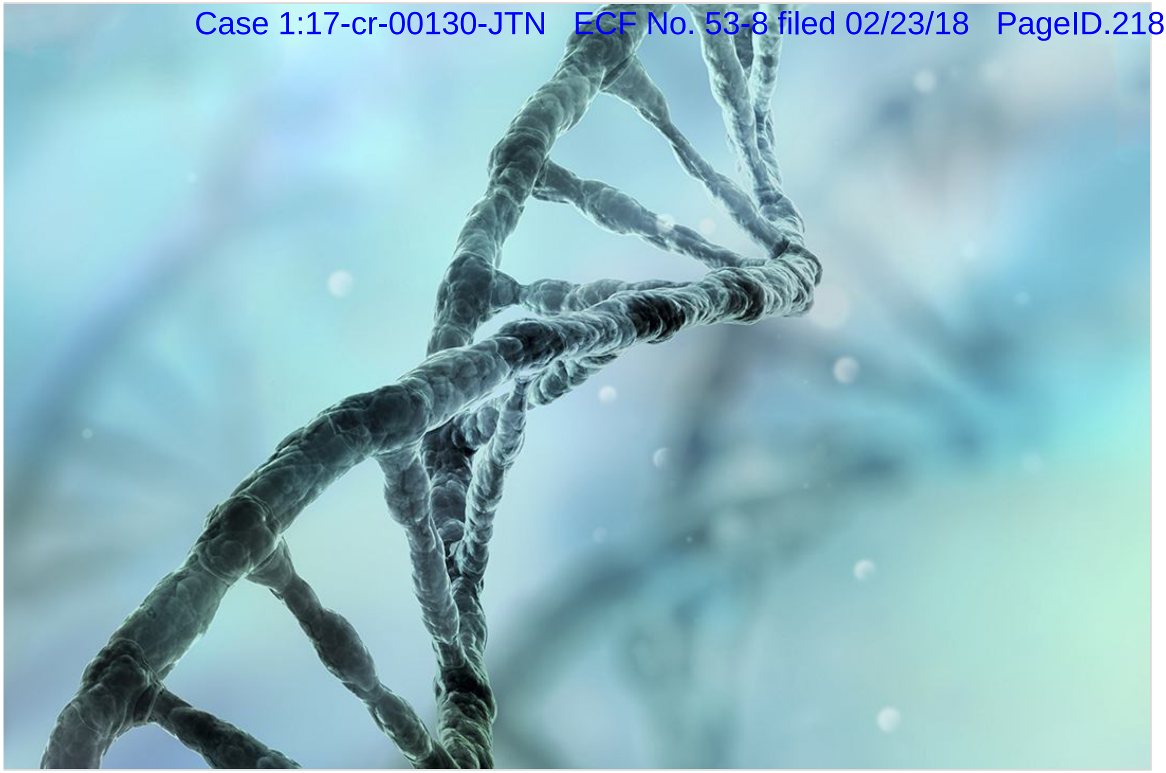
[131] *See* U.S. v. Tank, 200 F.3d 627 (9th Cir. 2000); Wisconsin v. Schroeder, 2000 WL 675942.

6 VAJLT 13

End of Document

Exhibit 8

Lael Henterly, The Troubling
Trial of Emanuel Fair



The Troubling Trial of Emanuel Fair

In a grisly murder case, the defense wants to know if the DNA evidence is reliable. They'll never find out.

By Lael Henterly

Wednesday, January 11, 2017 1:30am | NEWS & COMMENT

It was her first Halloween at the Valley View apartments in Redmond, and Arpana Jinaga was eager to celebrate. It was 2008, and Jinaga, an accomplished software developer, had recently moved to the area to work at a technology company. She decked out her apartment in Halloween decor and donned a red cape purchased earlier that day in anticipation of an evening spent with her neighbors.

The small, three-story complex was filled with revelers that night. Costumed residents and guests milled from one unit to the next, laughing, drinking, and admiring each other's costumes. A gregarious and gracious host, Jinaga welcomed several of her neighbors and their guests into her apartment, where they conversed and mingled before moving on to her neighbor's apartment, where the group partook of vodka shots. By 3 a.m., the festivities were winding down. Jinaga said goodbye to the remaining guests and headed back to her apartment.

That weekend Jinaga didn't leave her apartment or contact any of her friends or family. By Monday morning her parents, far away in India, had grown worried. Jinaga's father asked Jay, a family friend also living in the area, to stop by and check on his daughter. The door to her apartment swung open with a knock and, joined by one of Jinaga's neighbors, Jay ventured in. The men stepped cautiously into the apartment, and immediately smelled bleach. In the bathroom a blood-stained comforter filled the bathtub; in the bedroom, charred black satin sheets; and there, sprawled on the carpet beside her bed, was the 24-year-old programmer, bloody, naked, and soaked in bleach and motor oil.

Officers from the Redmond police department responded and quickly locked down the scene. The police collected samples from hundreds of items at the crime scene that could hold traces of the killer's DNA and sent them to the state crime lab for analysis. Meanwhile, Redmond detectives began trying to whittle down the list of suspects, which initially included a number of guests from the Halloween party.

Two years after the killing, the arduous investigation culminated in a charge of murder in the first degree with sexual motivation against Emanuel Fair, a friend of a woman who lived downstairs from Jinaga. Fair, who is also known as Anthony P. Parker, was at the Valley View apartments that night, one of a group of revelers who spent time with Jinaga in her apartment. Investigators believe that sometime between 3 a.m. and 8 p.m., Fair broke down Jinaga's door, raped, beat, and strangled her, then went to great lengths to try to scrub his DNA from the scene.

Whoever went to those great lengths, genetic material was left behind. Investigators were drawn to Fair because of a criminal history that includes a third-degree rape conviction, but it is the DNA that stands as the prosecution's strongest evidence.

Justice has not been swift. Six years after charges were filed, Fair, now 33, spends his days shuffling between the King County Superior Court and the county jail while his two defense attorneys fight for a fair trial, a representative jury, and the opportunity to cross-examine key witnesses. The trial is scheduled to begin Jan. 13, 2017, with prosecutors seeking a sentence of 45 years to life.

Pretrial motions filed by Fair's attorneys indicate that he received treatment different from that given other suspects at each step of the investigation. There is some merit to these claims. According to interrogation transcripts, for instance, the Redmond detectives cajoled the other chief suspect in the case, whereas they threatened Fair.

Then there is the DNA.

In recent decades, DNA evidence has become a ubiquitous crime-solving tool. The gold standard of forensic science, genetic profiles inferred from DNA evidence pave the way to exonerations, confessions, and convictions both on- and offscreen. But while single-donor DNA samples are straightforward to analyze, degraded samples that contain multiple people's DNA—such as those found at the scene of Jinaga's murder—are far more difficult to nail down. In fact, the DNA evidence that investigators are relying upon is of such low quality that a few years ago it would have been considered unreadable.

The case against Fair hinges on the forensic evidence, some matched by humans to a genetic profile, and some more complex mixtures analyzed by TrueAllele, a probabilistic genotyping software program that relies on sophisticated algorithms to analyze DNA mixtures that humans can't. The program runs genetic data through its 170,000 lines of code, rapidly inferring DNA profiles and delivering a likelihood ratio that indicates the odds of a match. Mark Perlin, CEO of Cybergenetics, the private company that developed TrueAllele, says the software is capable of analyzing mixture samples with six or more contributors.

Only some of the more complex mixtures of DNA were sent to the Cybergenetics lab. The match statistics delivered by the software were far more definite than the numbers the state crime lab had generated when they analyzed the same samples. For example, the WSPCL found that a DNA mixture on Jinaga's robe was 1,000 time more likely to contain Fair's DNA than that of an unrelated African American. TrueAllele found that same sample to be 56.8 million times more likely to include Fair's DNA.

Fair's defense attorneys sought access to TrueAllele's 170,000 lines of source code from Cybergenetics, but Perlin refused to make the code to his proprietary software available. Revealing TrueAllele's code, even just for review, Perlin told the court, would compromise the company's trade secrets, potentially causing irreversible commercial damage.

Fair isn't the first defendant to question Cybergenetics' computer program —defendants in six other states have petitioned unsuccessfully for access to their algorithmic accuser's code. Last year a man in New York was sentenced to 15 years in prison after the software found him to be one of four, five, or six individuals who had handled a handgun recovered in a park.

This is the first challenge of this type in Washington; previously, TrueAllele has been used only to exonerate, not convict. Perlin says that overall, TrueAllele has been used in 500 criminal cases since 2009.

DNA evidence carries a lot of weight with juries, but experts say that mixture DNA is far less reliable than DNA from a single donor. Most of our genetic code is identical, but in some spots on each strand there are alleles —variations that differ from one person to another. To determine a DNA match, investigators compare a DNA strand's 16 alleles; by the standards in most labs, 13 makes a match.

The system is not without its flaws. In 2010 researchers Greg Hampikian and Itiel Dror gave 17 expert forensic scientists the same DNA mixture and found that the results varied wildly from one scientist to the next. Those who were given information about the criminal case the DNA evidence would be used in were more likely to find evidence that implicated suspects.

Dror, a cognitive neuroscientist who worked on the study, says that it's a mistake to think that probabilistic genotyping software is able to objectively analyze DNA mixtures. Like a human, he says, the software has to make assumptions about what is and isn't useful information. With TrueAllele, no one but Perlin knows what those assumptions are. “Using software doesn’t solve the problem, because the human biases, assumptions, and discretions go into the software,” says Dror. “The software has human biases; to see what the biases are, we need to look at the software to see what it’s doing.”

Since TrueAllele doesn’t have to show its work, it would be difficult to discover if the code had errors. It’s not unfathomable that that could be the case. In 2015, investigators in Australia discovered an error in the code of a rival probabilistic genotyping program, STRMix, and announced that incorrect results were being used in 60 criminal cases, one a murder.

In September the President’s Council of Advisors on Science and Technology (PCAST) issued a report acknowledging that humans aren’t reliable interpreters of complex DNA mixtures and that computer software, while a step in the right direction, needs further scientific validation.

“The biggest issue is there is no truly independent assessment of TrueAllele or other programs,” says Amy Jeanguenat, CEO of the forensic consulting firm Mindgen. “They don’t work the same, and some are better at certain profiles and the community doesn’t know the benefits and weaknesses.”

There were other leads. Marc O’Leary, a convicted serial rapist and home invader who is serving a 327-year sentence in Colorado, was active in the area at the time, attacking women in their homes in a manner eerily similar to that of Jinaga’s killer; the detectives didn’t look into the similarities. Serial killer Israel Keyes visited the Seattle area that Halloween weekend, but when the FBI asked Seattle-area law enforcement if they knew of crimes that could have been committed by Keyes, the Redmond detectives didn’t respond.

Then there is the neighbor. Early in their investigation, detectives cast a wide net but focused much of their attention on the neighbor who discovered Jinaga’s body alongside Jay, going so far as to draft a probable-cause document supported by 48 pieces of evidence. (Since he has not been charged with a crime, *Seattle Weekly* is choosing to not publish the neighbor’s name.)

Jinaga's murder occurred around 8 a.m., but she was last seen around 3 a.m. when she retired to her apartment at the end of the evening. Another Valley View resident arrived home from work within five minutes of 3 a.m. that night and said he saw a man—who, unlike Fair, was not black—standing in Jinaga's doorway talking to someone inside. The suspected neighbor claimed not to recall making two calls to Jinaga on either side of 3 a.m. the night of the murder. Later that morning he went on an unplanned excursion to Canada, only to be turned around at the border. The man told Redmond detectives he had been "kind of wanting to explore."

In interviews with the police, that neighbor's friends and family members told detectives they believed it was possible that he had killed Jinaga––at least three people told detectives that he had expressed concern that he may have gone to her apartment in his sleep. The neighbor revealed to the detectives that he had been off his psychiatric medication at the time of the murder; the investigators neglected to follow up to determine what medication he was prescribed and what condition the drug was intended to treat.

The neighbor and Jinaga had been friends, but drifted apart when she got a motorcycle and started devoting more time to the PNW Riders, he explained to Redmond detective Brian Coats in an interview the day Jinaga's body was found. Their apartment manager, who told detectives she believed the neighbor may have killed Jinaga, recalled that he and Jinaga sometimes wrestled and that she also used to wrestle with him but stopped because he "doesn't stop, even when it's time to stop." Jinaga, who practiced mixed martial arts, also sometimes got carried away when the two were wrestling, the manager told detectives.

After Fair was charged, the manager backed away from her earlier statements, saying she wouldn't have accused the neighbor if she had known "there was this big black guy there."

In 2009 the Washington State Patrol Crime Lab analyzed the items gathered from the crime scene for DNA testing and came back with the profiles of three men. A mixed sample from Jinaga's neck was 3,803 times more likely to be a combination of Fair and Jinaga's DNA than of Jinaga and a random black man. Samples from two places on her robe were considered by the crime lab to be 120 and 1,000 times more likely to be from Fair and Jinaga than from Jinaga and a random black man. DNA that could belong to Fair or the neighbor was found in a mixture on the roll of tape. A motor-oil bottle purchased by Jinaga was found in a plastic bag, along with her robe in the dumpster; it contained a mixture that was 120 million times more likely to be Jinaga and the neighbor's DNA than that of Jinaga and an unknown individual. DNA from another Valley View resident was found on a bootlace that detectives believed served as the ligature in the murder. The man vehemently denied ever seeing the bootlace and detectives didn't pursue it further.

For more than a year, the Redmond detectives and the state patrol worked to identify Jinaga's killer. By May 2010, interview transcripts show that the Redmond detectives had whittled the suspect list down to two men, Fair and the neighbor.

Later that month the detectives and prosecutor met with the neighbor at the office of his criminal defense attorney, John Henry Browne. Detective Coats told the neighbor that Jinaga's dad and sister contacted him every couple of weeks to check the status of the case. The detectives needed to close the case; "just to bring them a little bit of closure is what our ultimate goal is in all of this," Coats told the neighbor. The detective tried to elicit information about the night of the murder that could eliminate his name from the suspect list, to no avail. The neighbor continued to assert that he didn't remember what happened the night of the murder.

In August 2010 Detective Coats approached the neighbor again, asking if he could do anything "to shake a memory loose that could help you be a better witness for us and not be such an asset for the defense."

"I really wish I had somethin' I could, but just don't—don't know," he replied. Transcripts reveal that the man continued to question whether he could have forgotten kicking in Jinaga's door the night of her murder.

Like the TrueAllele software, Jinaga's neighbor won't be available for cross-examination. Until recently the prosecution's case involved a single killer who broke into Jinaga's apartment and gagged, raped, and strangled her. More recently the prosecutor's office indicated that the next-door neighbor was not going to be available as a witness because their office wasn't going to grant him immunity, as they had in earlier interviews, and he couldn't testify without incriminating himself.

In a recent court filing, Senior Deputy Prosecuting Attorney Brian McDonald clarified his office's stance: The evidence implicating the neighbor in Jinaga's death doesn't exculpate Fair. McDonald noted that the neighbor may have been involved in the murder and continues to be a "person of interest." In another court filing, the state described him as an "uncharged accomplice."

Jinaga's friends are glad to hear someone is finally being tried for her death. "It's a long time coming—I didn't realize this guy was going up for trial," says Colt Bristow, a friend of Jinaga's from the PNW Riders. "Glad they're going forward with it."

news@seattleweekly.com

SEATTLEWEEKLY

Exhibit 9

Kelly, H., et al., A comparison of statistical models for the analysis of complex forensic DNA profiles

Science and Justice 54 (2014) 66–70

Contents lists available at ScienceDirect

Science and Justice

journal homepage: www.elsevier.com/locate/scijus

Emerging researcher article

A comparison of statistical models for the analysis of complex forensic DNA profiles

Hannah Kelly [a,b,*], Jo-Anne Bright [a], John S. Buckleton [a], James M. Curran [b]

[a] ESR, PB 92021, Auckland 1142, New Zealand
[b] Department of Statistics, University of Auckland, PB 92019, Auckland 1142, New Zealand

ARTICLE INFO

Article history:
Received 24 October 2012
Received in revised form 15 July 2013
Accepted 17 July 2013

Keywords:
Forensic DNA interpretation
Binary model
Drop model
Continuous model

ABSTRACT

Complex mixtures and LtDNA profiles are difficult to interpret. As yet there is no consensus within the forensic biology community as to how these profiles should be interpreted. This paper is a review of some of the current interpretation models, highlighting their weaknesses and strengths. It also discusses what a forensic biologist requires in an interpretation model and if this can be realistically executed under current justice systems.

1. Introduction

In forensic DNA analysis, a profile is typically produced from a biological sample collected from the scene of a crime and compared with the DNA profile of one or more persons of interest (POI). Traditional DNA analysis is sequential. Initially an electropherogram (epg) is produced. This raw output is processed by assigning peaks as allelic, stutter or artefactual. The deduced profile is then compared to the POI (if available), with the intention of producing either an inclusion, or an exclusion. If an inclusion is reached, then it is customary to provide a statistic to support the strength of the evidence. Analysis can involve either human or computerised processing, based on empirically devised guidelines, and can be complicated by factors such as the number of contributors to the profile, and the quality and quantity of the DNA.

Single source "pristine" profiles are relatively simple to interpret and their analysis has achieved worldwide acceptance as a reliable scientific method. However, profiles from crime scenes are frequently compromised in quality, or quantity, or both (LtDNA). Stochastic factors are often present in such compromised profiles. This complicates interpretation. These stochastic factors can include heterozygote imbalance, increased stutter peaks, allelic dropout, locus dropout, and drop in [1,2]. Complicating interpretation even further is that in many cases, crime scene samples contain DNA from two or more people. Such profiles are referred to as mixtures.

The interpretation of mixtures can be difficult. The number of contributors is often unclear. The presence of three or more alleles at any locus signals the existence of more than one contributor, although it often is difficult to tell whether the sample originated from two, three, or even more individuals because the various contributors may share alleles. The number of contributors to the mixture is often assigned either by using the fewest number of individuals needed to explain the alleles [3–5], or by maximum likelihood methods [6]. In many cases there will be a major and a minor contributor present in the sample and the profiles can be resolved and interpreted as single source profiles. However, many profiles cannot be separated and are deemed "unresolvable". These complex mixtures are challenging to interpret and as yet, there is no consensus as to how such profiles should be dealt with in the forensic biology community.

A 2010 article in New Scientist [7] highlights the disparity of practice in the interpretation of complex mixtures. In this article an epg from a previously analysed complex mixture was presented to 17 analysts in the same government laboratory for interpretation. Only one analyst agreed with the original finding, that the POI could not be excluded from the mixture. Four analysts deemed the evidence inconclusive, while the remaining 12 said that the POI could be excluded as having contributed to the mixture.

For a field which is widely regarded as objective, such a range of conclusions for the same evidence is worrying. Additionally, if the analyst is presented with the profile of a POI along with case circumstances strongly indicating that they are the offender, there is the perturbing issue of bias. If the accompanying statistic does not correctly represent the strength of the inclusion (or if no match statistic is provided) then there is the risk of the DNA evidence being misrepresented in court.

A 2005 study [8] highlights that not only are complex mixtures difficult to interpret, it can also be difficult to determine how many people have contributed to the mixture. The authors showed that more than 70% of four person mixtures could be wrongly interpreted as two or three person mixtures. In New Scientist [9] one of the authors from

[*] Corresponding author at: ESR Ltd, PB 92021 Auckland 1142, New Zealand. Tel.: +64 9 815 3670; fax: +64 9 849 6046.
E-mail address: Hannah.kelly@outlook.co.nz (H. Kelly).

1355-0306/$ – see front matter © 2014 Forensic Science Society. Published by Elsevier Ireland Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.scijus.2013.07.003

the 2005 study, Dan Krane, states that: “If you can't determine how many contributors there were, it is ludicrous to suggest that you can tease apart who those contributors were or what their DNA profiles were”.

The following work is a review of some of the current interpretation models. We attempt here to highlight the weaknesses and strengths of these models. We also attempt to address the question of what a forensic biologist requires in a model and if this can be realistically implemented under current justice systems.

2. Calculation of a statistical weight

The DNA Commission of the International Society of Forensic Genetics (ISFG) recommends the use of the likelihood ratio (*LR*) in mixture interpretation [10]. The *LR* is accepted to be the most powerful and relevant statistic used to calculate the weight of the DNA evidence. It is the ratio of the probability of the evidence (*E*) given each of two competing hypotheses, H_1 and H_2, given all the available information, *I*. The available information, *I*, is taken to include the knowledge of the genotypes of the known contributors, *K*, the POI, *S*, and any other relevant and admissible evidence:

$$LR = \frac{\Pr(E|H_1, I)}{\Pr(E|H_2, I)}.$$

The interpretation models discussed in this paper all utilise the likelihood ratio.

3. Interpretation models

3.1. The binary model

The binary model is probably better defined as a family of models rather than one specific model. The models in this family share the characteristic that they assign genotypes as possible or impossible given the data.

We define the genotype of the observed crime stain as *O*, and the genotypes of proposed donors as G_i for donor *i*. For an *N* donor mixture there are *N* proposed genotypes, G_i for each proposed combination. The *j*th combination in a set of *N* genotypes is denoted S_j. We can interpret the binary models as assigning a value of zero or one to $\Pr(O|S_j)$. The binary model assigns the values zero and one to the unknown probabilities, $\Pr(O|S_j)$, based on reasonable methods that approximate the relative values of $\Pr(O|S_j)$. In essence $\Pr(O|S_j)$ is assigned a value of zero if it is thought that this probability is very small relative to the other probabilities. $\Pr(O|S_j)$ is assigned a value of one if it is thought that this value is relatively large. As such, it is an approximation. Currently in most forensic biology laboratories this probability assignment is done manually and by the application of analysis thresholds and other rules based on empirical data.

Peak heights can vary between in epgs when replicates are run from the same sample. This variation between replicates from the same sample can be more dramatic if the sample is low template LtDNA. In LtDNA samples, some peaks at a locus may fail to reach the predetermined threshold to call a peak an allele in one replicate, but may exceed the threshold in a different replicate, therefore allowing it to be called. Since there is observable variation in replicates it is not possible that any crime scene profile (given a genotype set S_j) could occur with probability one, although zero is still possible. The reality is that all the probabilities, $\Pr(O|S_j)$, have some value in the interval [0,1).

The most rudimentary implementation of the binary model treats alleles as present or absent and does not take into account peak height information [11–13]. We will term this the qualitative binary model.

Consider a set of allelic peaks A_1..., A_M. All sets of *N* genotypes that have these *M* alleles and no others are deemed included. Genotype sets are constrained by H_1 and H_2 (termed the allowed sets). The *LR* is assigned using the ratio of the sum of the probabilities of all allowed sets under H_1 and H_2.

The computer programme POPSTATS, in common use in North America, implements this approach following the formulae of Weir et al. [12]. These formulae use the product rule and make no assessment of sampling uncertainty. This approach also appeared in the now obsolete DNAMIX I software [12]. It should be noted that this approach cannot be used if dropout is possible and if used may result in a seriously non-conservative assessment of the data. It is therefore not recommended for the interpretation of LtDNA or complex mixtures.

DNAMIX II extended this approach to include a subpopulation correction following NRC II recommendation 4.2 and implements the formulae of Curran et al. [13]. DNAMIX II makes no assessment of sampling uncertainty and, again, cannot be reliably used on profiles where dropout is possible.

DNAMIX III implements the formulae described in Curran et al. [13] and provides a limit on the confidence interval based on the work of Beecham and Weir [14]. The confidence interval itself is dependent on the extent of population substructure and the number of subpopulations. The software is not appropriate for profiles where dropout is possible.

Shortfalls in the qualitative binary approaches described above, such as the failure to take into account peak height and the inability to account for the possibility of dropout lead to the development of extensions which we will term the semi-quantitative binary model.

The semi-quantitative binary model declares some of the combinations that would have been allowed under the qualitative binary model as *possible* or *impossible* [5,15]. Scientists use expert judgement together with a number of empirical guidelines to decide which genotype combinations at a locus can be excluded [5]. This assignment is often based on expert judgement or heuristics employing limits on variation in the mixture proportion (*mx*) and heterozygote balance (*h*).

The semi-quantitative model is mainly applied manually. However, GeneMapper® *ID-X* is a programme designed for the automated designation of forensic STR profiles [16]. It incorporates a mixture analysis tool that uses the number of peaks, peak height information, *mx* and interpretation guidelines to resolve two person mixed profiles in a semi-automated fashion based on Gill et al. [3].

Traditionally, the semi-quantitative binary model accounts for the possibility of drop-out by omitting the locus or using the 2*p* rule. The 2*p* rule assigns the probability $2\Pr(A_i)$ for the observation of a single allele, A_i, whose partner may have dropped out. The 2*p* rule had been assumed to be conservative in all circumstances, however this has proved a false assumption and is no longer recommended for use [10,17].

One method to extend the binary model to profiles where dropout may have occurred (but alleles matching the POI are present within the profile) uses the ‘*F*’ designation to denote an allele that may have dropped out or ‘failed’. In this system the *F* designation represents any allele at the locus in question, including alleles already observed [18].

An alternate extension method uses a ‘*Q*’ designation in place of the *F*. A *Q* designation represents any allele at the locus except for those alleles already present. The formulae for the *Q* model can become very complex. As it is applied manually, this method is not readily extended to higher order mixtures (those containing more than two contributors) but there is the potential for automation of these extensions [19–21].

The UK Forensic Science Service (FSS) developed a software, PENDULUM, that is automated and applies rules based on empirical data to assist in designating genotype sets as possible or not possible and uses the *F* designation [15]. However, PENDULUM ends the process at these designations and does not proceed to calculate a *LR*, nor does it provide any other calculation of a statistical weight.

Binary models have served well for a number of years and in a great many cases, but with the advent of increasingly sensitive DNA analysis techniques, more samples containing low levels of DNA are now being submitted for analysis and these samples can often contain non-concordances.

The primary motivator for change is that the binary models described above cannot deal with a locus showing a non-concordance. This is a locus where at least one allele of the POI is not seen in the profile. In addition, none of the models can take into account multiple replicates. The challenges associated with the phenomena of dropout and drop in, in particular, have led to the evolution of a model which assesses the crime scene profile utilising primarily the concept of a probability of dropout.

3.2. The semi-continuous model

Fig. 1 shows two examples of non-concordances when the POI is the genotype (7,9). Example A shows a large concordant 7 peak which is just under the homozygote threshold and no peak at the 9 allele position. Example B shows a small concordant 7 peak and a below threshold 9 peak. Previously both examples would have been treated using the $2p$ rule under the binary models. If we use subjectivity to assess the two examples we can see that they are both quite different. In reality there is considerable support for the genotype 7,7 in example A while in example B there is more support for the genotype 7,9. Using the $2p$ rule in situations such as example A is non-conservative, which has led to the development of the Buckleton and Gill model [4,22].

The Buckleton and Gill model assigns a probability to the event of an allele not appearing, $\Pr(D)$. This is usually shortened to D (i.e. the probability that an allele would dropout) [19,22,23]. It can also factor in the presence of additional genetic material, referred to as drop in, $\Pr(C)$. In this model drop in is distinct from contamination. Drop in is not reproducible and is limited to only a few peaks per profile, whereas contamination refers to the presence of portions of reproducible extraneous DNA. This method also can cope with multiple replicates (for a more thorough discussion refer to Buckleton and Gill [24]). The probability of dropout appears in both the numerator and denominator of the LR. If the hypothesis gives information about the probability of dropout it will differ in the numerator and denominator. If the hypothesis gives no information then the probability of dropout is simply a property of the data, with a distribution related to a laboratory process.

The FSS implemented this approach in the software, LoComatioN [19]. However, the epg is still evaluated qualitatively first. The scientist must call peaks as alleles and assign stutter peaks. The assigned peaks are then entered into the computer programme and the probabilities of the profile for all possible genotype sets are calculated. The software can calculate a likelihood ratio for a range of propositions manually entered into the programme by the analyst. It enables a rapid evaluation of multiple propositions which would otherwise be laborious and error prone [19].

However, no peak height information is utilised when designating genotype sets. For example, in Fig. 2, all of the genotype combinations would be given the same weight [20,22,25].

Tvedebrink et al., [26,27] have suggested various improvements to the assignment of the probability of dropout. All of these methods use the profile itself to assess one or two covariates used to assign the probability of dropout. The treatment of the probability of dropout as a parameter assessed from the profile can be problematic as there is a recycling of the information. It would be better to treat the probability of dropout as a random variable and integrate it out [28]. This would require a sensible distribution to describe the probability of dropout. Such a distribution would vary from case to case. As yet such concepts have been mentioned but not implemented.

The semi-continuous model is an improvement in the way complex mixtures and LtDNA profiles are interpreted. However it still does not make full use of the available information from the epg. Consider the epg shown in Fig. 2. If we treat this as a two-person mixture, then six genotype combinations are deemed possible. These are:

Individual 1	Individual 2
7,9	11,13
7,11	9,13
7,13	9,11
9,11	7,13
9,13	7,11
11,13	7,9

The combinations 7, 11 : 9, 13 and 9, 13 : 7, 11 are well supported by the peak heights. However, under the semi continuous model (and binary models) the profile is assigned the same probability for all of the genotype combinations listed. When this concept is extended to multiple loci only one combination will be the most supported.

Addressing these shortcomings leads into the concept of the continuous model. This model seeks move away from very discreet all ($\Pr(O|S_j) = 1$) or nothing ($\Pr(O|S_j) = 0$) nature of the binary model by making better use of the available information.

3.3. Continuous models

We define a fully continuous model for DNA interpretation as one which assigns a value to the probability $\Pr(O|S_j)$ using some model for peak heights for all peaks in the profile. These models have the potential to handle any type of non-concordance and may assess any number of replicates without pre-processing and the consequential loss of information. Continuous models are likely to require models to describe the stochastic behaviour of peak heights and potentially stutter.

Many of the qualitative or subjective decisions that the scientist have traditionally handled such as the designation of peaks as alleles, the allocation of stutters and possible allelic combinations may be removed. Instead, the model takes the quantitative information from the epg such as peak heights, and uses this information to calculate the probability of the peak heights given all possible genotype combinations.

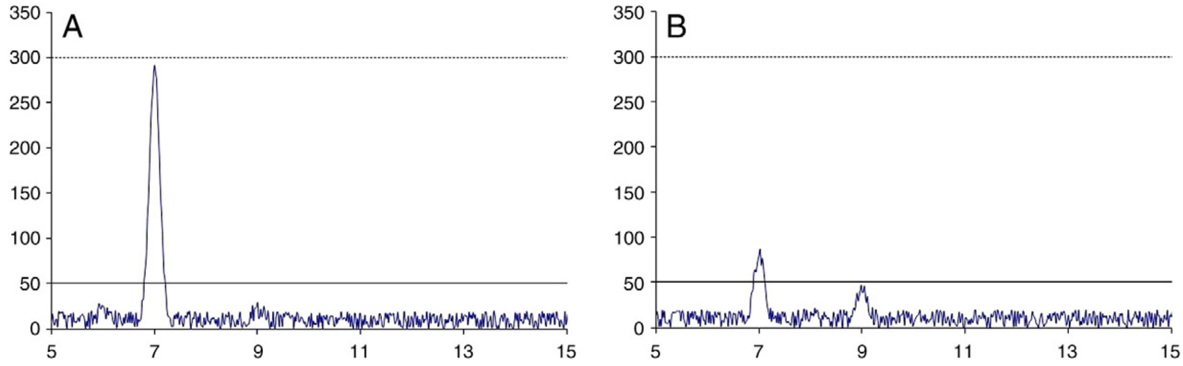


Fig. 1. Two examples of non-concordance where POI = 7,9. A large concordant 7 allele with no 9 peak observed (non-tolerable non-concordance) and B small concordant 7 allele with a non-concordant 9 peak visible sub-threshold (tolerable non-concordance). Stochastic threshold = 300 RFU, limit of detection 50 RFU.

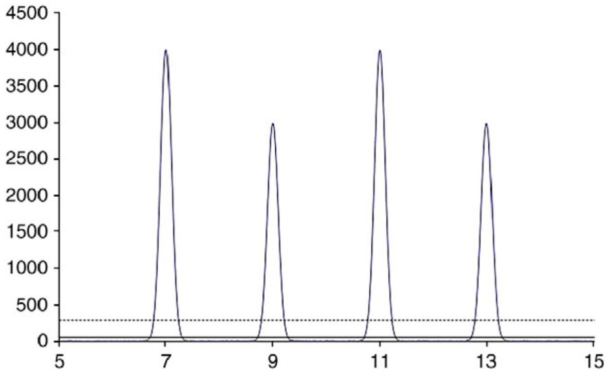


Fig. 2. Artificial epg of four-peak locus for a two-person mixture.

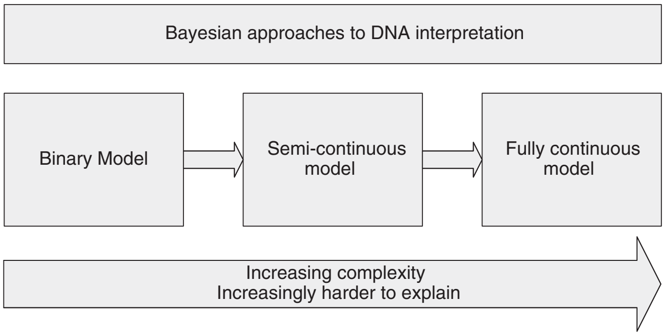


Fig. 3. Summary of the relationship of the different models for forensic DNA interpretation.

Removing the subjectivity or qualitative analysis of the profile will ensure consistency in DNA interpretation and reporting across laboratories.

TrueAllele is an example of commercial software implementing a continuous model [29].

4. General acceptance of a universal DNA model

It is appropriate, when assessing the advantages and weaknesses of these models, to begin by discussing which aspects of an interpretation model are desirable and/or suitable in the forensic context. Accuracy, reliability and comprehensibility are definitely desirable aspects of a DNA interpretation model. None of these are easy to define in this context.

If we think of the product of an interpretation model as a likelihood ratio, then we may think of accuracy as closest to the true answer. The true answer in DNA interpretation is somewhat elusive and plausibly does not exist at all. For this paper we will think of accuracy as making the best use of all the available information in a logically robust manner.

We will use the word reliability in this context to refer to the chance of serious misapplication of the method, either to a situation for which it is unsuited or misapplication to a situation for which it is suited.

Comprehensibility may come in two forms. Is the method comprehensible to the forensic scientist? Is the method explainable to a court? There is therefore interplay between comprehensibility to the scientist and reliability (Fig. 3). This point is possibly worth some expansion.

It is often assumed that complex and especially computerised methods are at most risk and this is plausible. However the risk exists for any method, computerised or otherwise, to be misunderstood.

The simplest model, when assessed against these criteria, is the qualitative binary model. One could easily justify the argument that it is the most comprehensible and reliable. And yet there is evidence that this is not so. This method is not suitable for profiles where dropout is possible but it is often applied to such profiles. It may be that if interpretation is not given sufficient importance within an organisation, then adequate training and research resources may not be invested in it. Organisations giving low priority to interpretation may choose simple systems and also have low investment in interpretation training and research. The conclusion is that even the simplest method may descend into the category of misunderstood.

The semi-quantitative binary model, when applied manually as is usually the case, is the one that has the scientist most intimately involved in the interpretation. This places considerable training and research requirements on the organisation but in many ways this is a good thing. Parameters of importance for interpretation need to be assessed such as variability in heterozygote balance and stutter peak heights. Staff must be trained to a high degree of competence but, again, this is desirable both from a professional standards viewpoint, and from the ability of the scientist to represent the evidence in court. However the binary model, in any version, is incapable of handling non-concordances. This is the primary motivator for a move away from this method. There is also the difficulty in extending the model to multi-person mixtures.

The Buckleton and Gill model retains many of the best aspects of the semi-quantitative binary model but allows extension to profiles showing non-concordances. Software is required to extend to mixtures of three or more persons or to multiple replicates. Programmes have been developed [19,22].

Coming finally to the continuous model; this approach is undoubtedly the premier choice in terms of accuracy as defined here, if we can adequately model the behaviour of peak height with empirical observation and verify the mathematical logic of the development from these foundations. Such methods will need to be consigned to a computer. Training and research demands will be considerable to underpin the approach and to allow scientists to represent the evidence in court [30]. Although the continuous model is unfamiliar to many forensic DNA scientists new ways of describing the method may facilitate education [31,32].

Additionally, computer software is only as reliable as the analyst that is using it. There is the risk that, with complicated automated programmes, analysts will not understand the limitations and the programme will be inadvertently used in situations where it is not appropriate to do so. However, properly developed and used, the continuous model will make the best use of the available information and give a considerable enhancement in objectivity [33,34]. Replicates may be easily accommodated [35,36]. The mathematics may be placed in the public domain by publication and hence it will be available for scrutiny by other qualified experts or subject to examination in court [37]. In many ways a well described mathematical process is more transparent than the often subjective decisions of experts.

5. Conclusion

DNA profiling is the stronghold in the characterisation of forensic biological evidence. The advent of increasingly more sensitive DNA analytical techniques has enabled scientists to generate profiles from samples that contain much lower amounts of DNA. This means that a wider range of evidence types can be analysed. However, the benefit of increased sensitivity, at times, means a reduction in profile quality and problems with profile interpretation due to the nature of the evidence types being sampled. Complex mixtures and LtDNA have stochastic factors present that complicate interpretation and current interpretation models are struggling.

Although extensions have been made to binary models we are being forced to move away from them, largely due to their inability to handle non-concordances but also by the difficulty in extending the semi-quantitative method to multi-person mixtures and the associated loss of information when expedients are used.

The options to move forward with are the semi-continuous Buckleton and Gill model and the continuous approach. Both of these are defensible scientifically. Of the two the continuous model makes the best use of the available information. Since both are likely to be encapsulated into software the risk of them being misused must be ameliorated. This will be a challenge but perhaps a worthwhile one in terms of professionalism.

What must be decided is if we should move towards a model that is most likely to deliver us the more accurate answer, yet the mechanics are complex to explain to a jury, or if we should move towards a method where the scientist has a more hands on approach and the model is easier to explain to a jury but does not use all the information available.

Realistically, the model which makes the best use of the available evidence has to be implemented. Therefore, we must advocate a move to a continuous method founded on sound biological models, which themselves are based on empirical data.

Acknowledgements

This work was supported in part by grant 2011-DN-BX-K541 from the US National Institute of Justice. Points of view in this document are those of the authors and do not necessarily represent the official position or policies of the U.S. Department of Justice. The authors would like to thank Johanna Veth, Pauline Simon, and two anonymous referees for their helpful contributions towards this paper.

References

[1] B. Caddy, G. Taylor, A. Linacre, A Review of the Science of Low Template DNA, Analysis, 2008.

[2] The Forensic Science Regulator, Response to Professor Brian Caddy's Review of the Science of Low Template DNA Analysis, , 2008.

[3] P. Gill, R.L. Sparkes, R. Pinchin, T. Clayton, J.P. Whitaker, J.S. Buckleton, Interpreting simple STR mixtures using allelic peak areas, Forensic Science International 91 (1998) 41–53.

[4] J.S. Buckleton, C.M. Triggs, S.J. Walsh, Forensic DNA Evidence Interpretation, CRC Press, Boca Raton, Florida, 2004.

[5] T. Clayton, J.P. Whitaker, R.L. Sparkes, P. Gill, Analysis and interpretation of mixed forensic stains using DNA STR profiling, Forensic Science International 91 (1998) 55–70.

[6] H. Haned, L. Pene, F. Sauvage, D. Pontier, The predictive value of the maximum likelihood estimator of the number of contributors to a DNA profile, Forensic Science International. Genetics 5 (2011) 281–284.

[7] L. Geddes, Fallible DNA evidence can mean prison or freedom, New Scientist (1971) 2774 (2010).

[8] D.R. Paoletti, T.E. Doom, C.M. Krane, M.L. Raymer, D.E. Krane, Empirical analysis of the STR profiles resulting from conceptual mixtures, Journal of Forensic Sciences 50 (2005) 1361–1366.

[9] L. Geddes, How DNA creates victims of chance, New Scientist 207 (2774) (2010) 8–10.

[10] P. Gill, C.H. Brenner, J.S. Buckleton, A. Carracedo, M. Krawczak, W.R. Mayr, N. Morling, M. Prinz, P.M. Schneider, B.S. Weir, DNA commission of the International Society of Forensic Genetics: recommendations on the interpretation of mixtures, Forensic Science International 160 (2006) 90–101.

[11] I.W. Evett, C. Buffery, G. Willott, D.A. Stoney, A guide to interpreting single locus profiles of DNA mixtures in forensic cases, Journal of the Forensic Science Society 31 (1991) 41–47.

[12] B.S. Weir, C.M. Triggs, L. Starling, L.I. Stowell, K.A.J. Walsh, J.S. Buckleton, Interpreting DNA mixtures, Journal of Forensic Sciences 42 (1997) 213–222.

[13] J.M. Curran, C.M. Triggs, J.S. Buckleton, B.S. Weir, Interpreting DNA mixtures in structured populations, Journal of Forensic Sciences 44 (1999) 987–995.

[14] G.W. Beecham, B.S. Weir, Confidence interval of the likelihood ratio associated with mixed stain DNA evidence, Journal of Forensic Sciences 56 (2011) S166–S171.

[15] M. Bill, P. Gill, J. Curran, T. Clayton, R. Pinchin, M. Healy, J. Buckleton, PENDULUM — a guideline based approach to the interpretation of STR mixtures, Forensic Science International 148 (2005) 181–189.

[16] Applied Biosystems, GeneMapper® ID-X Software Mixture Analysis Tool. Quick Reference Guide, http://www3.appliedbiosystems.com/cms/groups/applied_markets_marketing/documents/generaldocuments/cms_087508.pdf2008.

[17] J. Buckleton, C.M. Triggs, Is the 2p rule always conservative? Forensic Science International 159 (2006) 206–209.

[18] H. Kelly, J.-A. Bright, J. Curran, J. Buckleton, The interpretation of low level DNA mixtures, Forensic Science International. Genetics 6 (2011) 191–197.

[19] P. Gill, A. Kirkham, J. Curran, LoComatioN: a software tool for the analysis of low copy number DNA profiles, Forensic Science International 166 (2007) 128–138.

[20] H. Haned, Forensim: an open-source initiative for the evaluation of statistical methods in forensic genetics, Forensic Science International. Genetics 5 (2011) 265–268.

[21] D. Balding, likeLTD (likelihoods for low-template DNA profiles), https://sites.google.com/site/baldingstatisticalgenetics/software/likeltd-r-forensic-dna-r-code.

[22] D.J. Balding, J. Buckleton, Interpreting low template DNA profiles, Forensic Science International. Genetics 4 (2009) 1–10.

[23] P. Gill, J.P. Whitaker, C. Flaxman, N. Brown, J.S. Buckleton, An investigation of the rigor of interpretation rules for STR's derived from less that 100 pg of DNA, Forensic Science International 112 (2000) 17–40.

[24] J.S. Buckleton, P. Gill, Low copy number, Forensic DNA Evidence Interpretation, CRC Press, Boca Raton, 2004, pp. 275–297.

[25] P. Gill, L. Gusmão, H. Haned, W.R. Mayr, N. Morling, W. Parson, L. Prieto, M. Prinz, H. Schneider, P.M. Schneider, B.S. Weir, DNA commission of the International Society of Forensic Genetics: Recommendations on the evaluation of STR typing results that may include drop-out and/or drop-in using probabilistic methods, Forensic Science International. Genetics 6 (2012) 679–688.

[26] T. Tvedebrink, P.S. Eriksen, H.S. Mogensen, N. Morling, Estimating the probability of allelic drop-out of STR alleles in forensic genetics, Forensic Science International. Genetics 3 (2009) 222–226.

[27] T. Tvedebrink, P.S. Eriksen, H.S. Mogensen, N. Morling, Statistical model for degraded DNA samples and adjusted probabilities for allelic drop-out, Forensic Science International. Genetics 6 (2012) 97–101.

[28] D. Balding, Integrating *D*, Personal Communication, , 2010.

[29] M.W. Perlin, M.M. Legler, C.E. Spencer, J.L. Smith, W.P. Allan, J.L. Belrose, B.W. Duceman, Validating TrueAllele® DNA mixture interpretation, Journal of Forensic Sciences 56 (2011) 1430–1447.

[30] M.W. Perlin, Easy reporting of hard DNA: computer comfort in the courtroom, Forensic Magazine 9 (2012) 32–37.

[31] Cybergenetics, DNA identification for scientists, http://www.cyben.com/information/courses/page.shtml2012.

[32] M.W. Perlin, DNA identification science, in: C.H. Wecht (Ed.), Forensic Sciences, LexisNexis Matthew Bender, Albany, 2012.

[33] M.W. Perlin, A. Sinelnikov, An information gap in DNA evidence interpretation, PLoS One 4 (2009) e8327.

[34] J. Curran, A MCMC method for resolving two person mixtures, Science & Justice 48 (2008) 168–177.

[35] V.L. Pascali, S. Merigioli, Joint Bayesian analysis of forensic mixtures, Forensic Science International. Genetics 6 (2012) 735–748.

[36] J. Ballantyne, E.K. Hanson, M.W. Perlin, DNA mixture genotyping by probabilistic computer interpretation of binomially-sampled laser captured cell populations: combining quantitative data for greater identification information, Science & Justice 53 (2) (2013) 103–114.

[37] M.W. Perlin, J. Galloway, Computer DNA evidence interpretation in the Real IRA Massereene terrorist attack, Evidence Technology Magazine 10 (2012) 20–23.

Exhibit 10

Seth Augustine, DNA Mixtures Topic of ISHI Talks, NIST Testing – and Conflict of Interest Accusations

DNA Mixtures Topic of ISHI Talks, NIST Testing—and 'Conflict of Interest' Accusations

10/05/2017 - 1:19pm 1 Comment by Seth Augenstein

Seth Augenstein
Senior Science Writer
@SethAugenstein
Full Bio >

DNA

DNA mixtures are a growing concern in the world of forensic science. The detection of even miniscule touch DNA in a pool of blood, or the skin cells left behind on the handle of a kitchen knife that later becomes a murder weapon, is now more possible than ever before with hypersensitive instruments. Forensic experts contend detectives and scientists need to be able to understand how the invisible ATCG alphabet soup at a crime scene may implicate a murderer, an innocent person—or both.

Forensic science has been grappling with this, and software solutions involving complex statistical modeling and big-data computing have emerged to replace previously subjective expert analysis. DNA mixtures comparison was a major topic of presentations yesterday at the annual International Symposium for Human Identification, with the industry leader STRmix the focus of several talks. The forensic analysis of complex mixtures will be a major research project, the National Institute for Standards and Technology also announced Tuesday. But the stiff competition between the two major technologies—STRmix and TrueAllele—has led to accusations that "conflicts of interest" at the government agency favor one of the tools, and are therefore driving the science.

STRmix, Validated by the FBI

Several ISHI presentations focused on STRmix, a technology developed by scientists in New Zealand, where it has been in use for routine casework since 2012.

An FBI analyst presented at the Seattle meeting Wednesday morning, showing how the Bureau had completed internal validation of the tool. That validation involved assessing more than 2,800 mixtures of 3 to 5 different people, using the tool at 31 different laboratories nationwide, according to Tamyra Moretti of the FBI.

Another analyst showed how STRmix statistics have cracked criminal cases. Cristina Rentas of the Florida-based DNA Labs International presented several cases, one of which showed how STRmix comparisons actually changed the direction of detectives' suspicions. A three-person mixture was from a weapon, and the genetic material showed the victim was almost certainly in there. But the suspect was a different story. First, the STRmix analysis showed the suspect was almost assuredly part of the mixture (12 million times more likely). But the second run was distinctly less certain (only 17 times more likely). What had changed were the assumptions programmed into the software at the beginning of the tests, Rentas said. Once they started doing more genetic testing, adding more data from people surrounding the case, they hit on the shocking result: the main suspect’s brother was actually the culprit.

“It was head-spinning for us—it didn’t make sense to us, but then we decided to break down the likelihood ratios,” said Rentas. “As more and more labs are starting to bring online STRmix, and more and more labs are starting to use probabilistic genotyping, it’s important to share that this is something that can make a difference in the case—and bring a case in a new direction.”

STRmix is currently in 29 laboratories across the U.S., as well as in places like Australia, according to John Buckleton of the New Zealand-based Institute for Environmental Science and Research, the creator of STRmix. In the U.S., the software is gaining traction, coming online in dozens more laboratories across the nation, the company recently announced. The Bureau of Alcohol, Tobacco, Firearms and Explosives (ATF) announced last month it would use it, as would a new handful of sheriff’s offices in California, Florida and Illinois.

Buckleton has pointed to recent cases like one in Michigan where STRmix solidified a prosecutor’s case against a rapist based off a DNA mixture on the victim’s toes.

“I am excited at the growth STRmix has experienced in the U.S. The U.S. would now be considered a leader in the field of probabilistic genotyping,” Buckleton told *Forensic Magazine* this week.

The three developers of the software from Australia and New Zealand “do not receive any benefit direct or indirect from sales of STRmix” because they are civil servants paid by their governments, according to Buckleton’s website.

NIST to Assess

The DNA mixture question has been the subject of critical reports such as the watershed 2009 National Academy of Sciences report “Strengthening Forensic Science,” and last fall’s President’s Council of Advisors on Science and Technology, or PCAST, report on forensic science. That PCAST report recommended a complete reevaluation of DNA mixture analysis—including STRmix and TrueAllele.

NIST will now assess the reliability of picking apart DNA mixtures from multiple contributors. The "scientific foundation review" will be led by NIST Fellow John Butler, a noted DNA expert, and Sheila Willis, the former director of Ireland's national forensic laboratory, who is a chemist.

"The goal is not to undermine these methods, but to determine their bounds of reliability so they can be used appropriately," said Butler, in a NIST statement.

NIST did not specify which DNA mixture tools will be analyzed in their announcement.

Buckleton had served as a guest lecturer and researcher at NIST for roughly two years, until earlier this year—after his role there was questioned by the competitor, TrueAllele.

TrueAllele—and Conflicts of Interest Alleged

The major competitor to STRmix is TrueAllele, a software made by the Pittsburgh-based Cybergenetics. The software was created by Mark Perlin, who heads the company, and is a former Carnegie Mellon scientist.

TrueAllele is currently in at least seven labs across the country, and another seven have access to the system, Perlin told *Forensic Magazine*. (TrueAllele was used by the New York State Police crime lab for several years. But the lab was later enmeshed in allegations and counter-allegations involving accreditation, and a lawsuit alleging civil rights violations.)

TrueAllele has been used in more than 40 U.S. states, by both prosecutors and defense teams. Two men who served years in the Indiana prison system for a 1989 gang rape were exonerated early this year after TrueAllele proved they could not have been in the DNA mixtures collected.

Perlin says his software is superior with more complex mixtures. But he alleges his technology has suffered from "conflict of interest" at NIST—where Buckleton, the STRmix creator, was a guest lecturer up until last year.

Last year, Perlin sent a letter to the U.S. Department of Commerce asking for a look into the "special status" of

Buckleton, ESR and STRmix.

Buckleton's place within NIST had given STRmix an unfair advantage, despite the scientific considerations, Perlin alleged in the July 8, 2016 letter.

"The closeness between NIST and ESR is apparent to the forensic DNA community," he wrote. "This special relationship between your agency and a large foreign company may have unleveled the playing field for a small American innovator.

"The impact goes far beyond commerce, since widespread usage of weak crime-fighting DNA technology affects justice for all Americans," he adds.

Within months of that letter, Buckleton was no longer at NIST.

Earlier this year, NIST referred *Forensic Magazine* to the Department of Commerce for comment on the conflict allegations, and Buckleton's departure. The Department of Commerce at first said a response was forthcoming, and did not respond to months of further emails from *Forensic Magazine*.

Buckleton, this week, said he did not want to delve into the specifics of his departure from NIST.

"Suffice to say that I enjoyed the two years I spent at NIST and was sorry to leave," he wrote to *Forensic Magazine*. "My departure, though, led me to work with Professor Bruce Weir with the Department of Biostatistics and the Director of the Institute for Public Health Genetics at the University of Washington Seattle. Professor Weir is a leader in the field of evidence interpretation, so working with him has been tremendously exciting."

The acting NIST director sent Perlin and TrueAllele an April 2017 letter that indicated the agency wouldn't necessarily favor STRmix.

"Commercials products, materials, and instruments are identified in NIST's communications and documents for the sole purpose of adequately describing experimental or test procedures," wrote Kent Rochford, the acting NIST director. "In no event does such identification imply recommendation or endorsement by NIST of a particular product; nor does it imply that a named material or instrument is necessarily the best available for the purpose it serves.

"NIST strives to avoid even the appearance of endorsing commercial products, while working with a breadth of measurement techniques," Rochford added.

Perlin told *Forensic Magazine* this week that NIST shouldn't be conducting the tests on DNA mixture software at all.

"A biased referee cannot conduct a fair study," he said.

Tests at a New York Murder Trial

The two programs were put to the test in a high-profile murder trial in Upstate New York last year. Oral Nicholas Hillary stood accused of strangling the 12-year-old son of his ex-girlfriend, in a case with racial undertones within a small community near the Canadian border.

TrueAllele was first consulted on the case. Every sample distanced Hillary from the scene—including the fingernail scrapings, the company said. That evidence was not used in court, however.

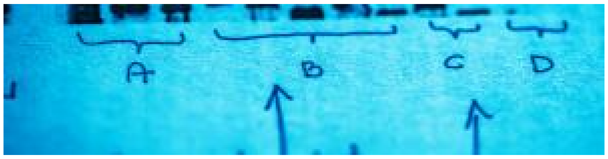
STRmix was enlisted by prosecutors next. Buckleton told the court he had assessed the complex genetic mixture under the boy's fingernails—and determined Hillary was in there.

However, the judge disagreed with the stated methodology in his decision, ruling he would not allow Buckleton to "pick and choose data from different 'reliable sources,'" for the prosecution's case, since the science had not been proven within the state's crime labs.

"Here, the lack of internal validation by the New York State Police crime lab, as candidly admitted by Dr. Buckleton, precludes use of the STRmix results," the judge ruled. "Neither the STRmix nor the (Random Match Probability) results may be used in this case."

Hillary walked free, with many believing him innocent, and other onlookers convinced he was guilty.

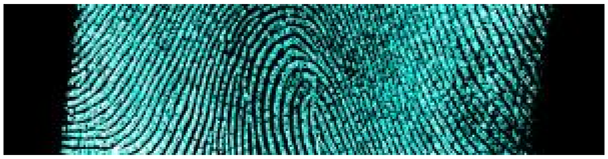
RELATED READS



Subjective DNA Mixture Analysis, Used in Thousands of Cases, Blasted by WH Panel

Austin Crime Lab Shut Down Over 'Concerns'

ISHI 2017: Rapid DNA, Genetic Mixtures, Next-gen and the Cold Trail of Dead Serial Killers



Do You Have What It Takes to Be a Forensic Fingerprint Examiner?

1 Comment **Forensic Magazine** 1 **Login**

♡ **Recommend** **Share** **Sort by Best**

Join the discussion…

LOG IN WITH

OR SIGN UP WITH DISQUS ?

Name

DNA Analyst • 4 months ago

Great article. The article also exposes the problem of commercial interests over good science. Both commercial providers with their marketing battle and "in fighting" are thwarting the general acceptance of Probabilistic Genotyping (PG) in the U.S. However, there is another group, a collective of several dozen labs known as the eDNA Consortium that has implemented a Continuous PG model with advanced mixture deconvolution capabilities. This PG software is internally named BulletProof and implements the best of EuroForGen's (EFG) algorithms known as EuroForMix. EFG is the European equivalent to the U.S. eDNA Consortium—neither are a commercial venture. EFM was developed by Øyvind Bleka, Geir Storvic, and Peter Gill and enhance by the eDNA Consortium to support high throughput casework.

BulletProof is active within the consortium and currently being used coast to coast. Admissibility challenges are easily overcome since they are not a commercial interest and freely share the source code and all underlying algorithms. In other words, it does not suffer from the "black Box" syndrome. Anyone interested in a robust PG software, that rivals the capabilities of the commercial solutions, just reach out to the eDNA Consortium and they'll

Exhibit 11

NIST to Assess the Reliability of Forensic Methods for Analyzing DNA Mixtures www.nist.gov/news-events/news/2017/10/nist-assess-reliability-forensic-methods-analyzing-dna-mixtures

National Institute of Standards and Technology
U.S. Department of Commerce

NIST to Assess the Reliability of Forensic Methods for Analyzing DNA Mixtures

October 03, 2017

A DNA technician prepares a sample for analysis.

Credit: Fred W. Baker/Department of Defense

The National Institute of Standards and Technology (NIST) will undertake a study to assess the reliability of forensic methods for analyzing DNA evidence that, if misapplied, could lead to innocent people being wrongly convicted. The study will focus on DNA mixtures involving three or more people, and on very small quantities of DNA also known as touch DNA.

The process of interpreting DNA mixtures and touch DNA can be subjective, and there are currently no clearly defined standards for doing so. The result: different analysts might come to different conclusions given the same evidence.

NIST Fellow and forensic DNA expert John Butler (https://www.nist.gov/people/john-butler) will lead an interdisciplinary team of scientists in conducting the study. The team includes Sheila Willis, the former director general of Ireland's national forensic laboratory, Forensic Science Ireland (FSI), who NIST recently brought on to work on this project. The researchers will also seek input from outside experts.

Many people consider DNA analysis to be an especially reliable forensic method, and often, it is. If blood, semen or other biological evidence is left at a crime scene, forensic scientists can use it to produce a DNA profile—a sort of genetic fingerprint—that can reliably identify a suspect.

Rigorous scientific studies have shown that when the evidence contains DNA from only one or two people, DNA profiles are extremely reliable. But when the evidence includes a mixture of DNA from three or more people, it can be difficult to tease apart the different profiles, or in some cases, to even determine how many profiles are present.

In addition, DNA methods have gotten so sensitive that investigators no longer need a blood or semen stain to generate a DNA profile. Today, labs will sometimes attempt to generate a DNA profile from, for instance, a few skin cells left behind when someone touched something at a crime scene (thus the name, "touch DNA"). But when analysts dial up the sensitivity of their methods, the data can end up including meaningless information, or "noise," that makes it difficult to interpret.

Many police agencies now routinely swab doorknobs and other surfaces for touch DNA when investigating property crimes. But if many people have touched those surfaces, the result may be a complex, low-level DNA mixture that is difficult, or impossible, to interpret reliably.

"Some labs won't do anything with that kind of evidence," said Butler. "Other labs will go too far in trying to interpret it."

The goal of the study is to measure the reliability of DNA profiling methods when used with different types of DNA evidence such as two-person mixtures versus four-person mixtures, and with different quantities of touch DNA. Crime laboratories, the courts and other institutions can then use information from this study to decide which methods pass muster.

"The goal is not to undermine these methods, but to determine their bounds of reliability so they can be used appropriately," Butler said.

While Willis is not a DNA expert, she is a chemist with deep experience in laboratory management. She said the study will be about more than just DNA technology. "We'll also be looking at what quality control systems and training are needed to support the people who collect and analyze the evidence."

NIST is calling this study a "scientific foundation review," and it will be designed in part on recommendations in a landmark 2009 report from the National Academy of Sciences (https://www.nap.edu/catalog/12589/strengthening-forensic-science-in-the-united-states-a-path-forward), which called for studies that demonstrate the validity of forensic methods.

While NIST's first scientific foundation review will look at DNA methods, the researchers aim to build a framework that will be extendable to other forensic methods. NIST intends to begin a scientific foundation review of bite mark evidence later this year.

This research is part of a larger effort by NIST to strengthen forensic science (https://www.nist.gov/topics/forensic-science) through research and the promotion of technically sound forensic science standards.

Media Contact

Rich Press (https://www.nist.gov/people/rich-press)

richard.press@nist.gov (https://www.nist.govmailto:richard.press@nist.gov)
(301) 975-0501 (https://www.nist.govtel:(301) 975-0501)

Related News

- NIST Research Enables Enhanced DNA "Fingerprints" (https://www.nist.gov/news-events/news/2016/12/nist-research-enables-enhanced-dna-fingerprints)
- Scientists Automate Key Step in Forensic Fingerprint Analysis (https://www.nist.gov/news-events/news/2017/08/scientists-automate-key-step-forensic-fingerprint-analysis)

Related Links

- Forensic Science (https://www.nist.gov/topics/forensic-science)

RELATED CONTENT

- How to Quantify the Weight of Forensic Evidence: A Lively Debate (https://www.nist.gov/news-events/news/2016/06/how-quantify-weight-forensic-evidence-lively-debate)
- NIST Research Enables Enhanced DNA “Fingerprints” (https://www.nist.gov/news-events/news/2016/12/nist-research-enables-enhanced-dna-fingerprints)
- New SRM Allows Accurate Accounting of Forensic DNA (https://www.nist.gov/news-events/news/2007/10/new-srm-allows-accurate-accounting-forensic-dna)

Created October 03, 2017, Updated November 27, 2017

Exhibit 12

William C. Thompson,
Are Juries Competent to
Evaluate Statistical Evidence?

ARE JURIES COMPETENT TO EVALUATE STATISTICAL EVIDENCE?

WILLIAM C. THOMPSON*

I

INTRODUCTION

The issue of jury competence has arisen due to concerns about the ability of jurors to deal appropriately with the increasing complexity of evidence presented in trials. While most attention has focused on complex civil litigation,[1] criminal trials have grown more complex as well, due in part to revolutionary advances in the forensic sciences.[2] New procedures for criminal identification, such as protein gel electrophoresis, DNA typing, gas chromotography, and neutron activation analysis have recently become available for use at trial.[3] Additionally, the technology behind many of the more traditional identification techniques, such as bite mark comparison and hair comparison, has advanced in recent years.[4] For a juror to understand and evaluate the technology underlying these techniques is often a formidable task in itself.

Adding further to the difficulty is the probabilistic nature of much of this new evidence. The results of forensic tests are often meaningful only if they are accompanied by statistical data. For example, evidence that the defendant in a rape case has genetic markers matching those in semen recovered from a rape victim cannot be evaluated without statistical information on the frequency of the matching markers in the population. Because forensic tests are often less than perfectly reliable, statistical data on the error rate of the test may be necessary as well.[5] Hence, jurors may hear that a criminalist compared a sample of the defendant's blood to a semen sample taken from the rape victim using a procedure known as protein gel electrophoresis and found that the two samples contain a common set of genetic markers that collectively occur in only 1.5 percent of the population. The jurors may also hear, however, that proficiency tests have found that criminalists misclassify

* Associate Professor, Program of Social Ecology, University of California, Irvine.

[1] *See, e.g.*, M. SAKS & R. VAN DUIZEND, THE USE OF SCIENTIFIC EVIDENCE IN LITIGATION (1983); Austin, *Jury Perceptions on Advocacy: A Case Study*, 8 LITIGATION 15 (1982); Lempert, *Civil Juries and Complex Cases: Let's Not Rush to Judgment*, 80 MICH. L. REV. 68 (1981).

[2] P. GIANNELLI & E. IMWINKELRIED, SCIENTIFIC EVIDENCE at xxi (1987).

[3] *Id.*

[4] *Id.* at 369-83, 1013-39.

[5] Among forensic scientists there is a "growing recognition that in many cases the results obtained yield their maximum information only if statistical methods and calculations of probability are used." Walls, *Ten Years of Forensic Science—1964-73*, 1974 CRIM. L. REV. 504, 505.

genetic markers in blood in semen at rates ranging from 1 to 6 percent per marker. What should jurors make of such evidence? What do they make of it?

The use of such statistical data in court is growing rapidly.[6] According to one authority, "our criminal justice system is now at the threshold of an explosion in the presentation of mathematical testimony."[7] The complexity of such testimony has raised concerns about the ability of jurors to deal with such evidence appropriately.[8] Empirical studies of the ability of lay individuals to use the type of statistical evidence presented in criminal trials have emerged only recently.[9] While this new literature is small and full of gaps, it is beginning to define the strengths and weaknesses of statistical reasoning by laypersons in ways that should prove quite helpful to courts facing decisions about the admissibility of statistical evidence and about the manner in which statistics should be presented to the jury.

II

THE NATURE OF STATISTICAL EVIDENCE IN CRIMINAL TRIALS

A discussion of jurors' competence to evaluate statistical evidence must necessarily begin with a description of the statistics jurors may encounter in a criminal trial and with a discussion of how jurors should evaluate those statistics.[10] There are two basic types of statistics: base rates and error rates. A base rate measures the frequency at which an event or characteristic occurs in a population.[11] An error rate measures the frequency at which a test or procedure produces wrong results. Although an error rate is a type of base rate, error rate statistics raise special issues distinct from those surrounding

[6] E. IMWINKELRIED, THE METHOD OF ATTACKING SCIENTIFIC EVIDENCE (1982); Jonakait, *When Blood is Their Argument: Probabilities in Criminal Cases, Genetic Markers, and Once Again, Bayes' Theorem*, U. ILL. L. REV. 369 (1983); Walls, *supra* note 5, at 505.

[7] Jonakait, *supra* note 6, at 369.

[8] Saks & Kidd, *Human Information Processing and Adjudication: Trial by Heuristics*, 15 LAW & SOC'Y REV. 123 (1980); Tribe, *Trial by Mathematics: Precision and Ritual in the Legal Process*, 84 HARV. L. REV. 1329 (1971).

[9] *See, e.g.*, J. GOODMAN, PROBABILISTIC SCIENTIFIC EVIDENCE: JURORS' INFERENCES (1988); Faigman & Baglioni, *Bayes' Theorem in the Trial Process*, 12 LAW & HUM. BEHAV. 1 (1988); Thompson, Britton & Schumann, *Jurors' Sensitivity to Variations in Statistical Evidence*, J. APPLIED SOC. PSYCHOLOGY (in press); Thompson & Schumann, *Interpretation of Statistical Evidence in Criminal Trials: The Prosecutor's Fallacy and the Defense Attorney's Fallacy*, 11 LAW & HUM. BEHAV. 167 (1987); J. Goodman, Jurors' Comprehension of Scientific Evidence (June 1988) (paper presented at the Meeting of the Law & Society Association, Vail, Colo.); E. Schumann & W. Thompson, Effects of Attorney's Arguments on Jurors' Interpretation of Statistical Evidence (June 1989) (paper presented at the Meeting of the Law & Society Association, Madison, Wis.); W. Thompson, J. Meeker & L. Britton, Recognizing Conditional Dependencies in Evidence: Effects of Group Deliberation (June 1987) (paper presented at the Meeting of the Law & Society Association, Washington, D.C.).

[10] The information reported here was gleaned from a review of appellate opinions and from a recent survey in which fifty forensic scientists in California were questioned about the types of statistical evidence they present in court and the ways in which they present it. N. Miller, The Role of Statistical Scientific Evidence in Criminal Trials: A Survey of Criminologists (1986) (unpublished thesis, University of California, Irvine). A cogent discussion of the various types of statistical evidence is also provided by Kaye, *The Admissibility of "Probability Evidence" in Criminal Trials—Part II*, 27 JURIMETRICS J. 160, 161-64 (1987).

[11] *See generally*, Koehler & Shaviro, *Veridical Verdicts: Increasing Verdict Accuracy Through the Use of Overtly Probabilistic Evidence and Methods*, 75 CORNELL L. REV. (in press).

other statistics on the frequency of events. Accordingly, base rates and error rates will be discussed separately.

A. Base Rates

Base rate statistics are usually developed from empirical studies in which a population, or a sample drawn from the population, is surveyed to determine the frequency of the event or attribute. For example, a survey showing 40 percent of a sample of Caucasians have type A blood establishes a base rate of type A blood among Caucasians. A study showing that 15 percent of the taxicabs in a city are green establishes a base rate of green cabs in the city. A study showing that 90 percent of defendants tried for burglary are convicted establishes a base rate of convictions among burglary defendants. The base rate of an event or attribute is equal to the probability that it will be present in a randomly selected member of the relevant population.[12]

1. *Directly Relevant and Indirectly Relevant Base Rates.* Base rate statistics can be used to prove a fact in two distinct ways. In some instances, the base rate is directly relevant to a target outcome because it directly expresses the frequency of that outcome. When a pedestrian is struck by a bus of unknown origin, evidence that a particular company operated 90 percent of the buses on that route is directly relevant to the question of who owned the bus. Similarly, when a defendant possessing heroin has been charged with concealing an illegally imported narcotic, evidence that 98 percent of all heroin is illegally imported is directly relevant to the question of whether the heroin possessed by the defendant was illegally imported. In such instances, the base rate is said to establish a prior probability of the target outcome.[13] If 90 percent of the buses that could have been involved in the accident are owned by a particular company, then there is a prior probability of .90 that that company owned the offending bus.

In other instances, the base rate is only indirectly relevant to a target outcome, and must be combined with other information before any probabilistic assessment of the target outcome is possible. When forensic tests link a criminal defendant to a crime by showing his blood type matches that of the perpetrator, evidence that the blood type is found in 5 percent of the population is relevant to the ultimate issue of the defendant's guilt, but only indirectly. The base rate of the blood type does not, by itself, reveal anything about the likelihood of the target outcome—the defendant's guilt—and thus, unlike a directly relevant base rate, does not establish a prior probability of the target outcome. Instead, it speaks to the likelihood the defendant might, by chance, have a "matching" blood type if innocent, and thus helps to establish the value of the forensic evidence.

[12] *Id.*

[13] The prior probability of a target outcome is the probability a reasonable person would assign to that outcome prior to receiving any case-specific or individuating information.

The use of "directly relevant" base rates as evidence in court has been controversial, particularly where the base rate is the sole evidence of a target outcome. Base rates of this sort have been labeled "naked statistical evidence"[14] and have generally been held inadmissible.[15] A few courts, however, have admitted such evidence.[16] The most widely discussed case involving "naked statistical evidence" is *Smith v. Rapid Transit,*[17] in which plaintiff was struck by a hit-and-run bus and based her claim that the bus was the defendant's solely on evidence that the defendant operated 90 percent of the buses in the city. The Massachusetts Supreme Judicial Court sustained defendant's motion for summary judgment on grounds that the base rate statistic was insufficient to make a case against the defendant in the absence of more particularized proof of the ownership of the offending bus.

While most commentators agree with this holding, they disagree on the rationale. One group, which has been labeled "anti-Bayesian,"[18] argues that base rates are inherently inferior to more particularized evidence and have little or no relevance unless they reflect a "suitably narrowed down reference class."[19] By this account, the frequency of defendant's buses among all buses in the city is merely "background information" that does not necessarily reflect the likelihood that the hit-and-run bus was defendant's, and therefore is an insufficient basis for a holding in plaintiff's favor.[20] Other commentators, sometimes labeled Bayesians, maintain that base rates need not meet any standard of specificity in order to be relevant.[21] By their account, the fact that defendant operates 90 percent of the buses in the city is highly relevant because it establishes a prior probability of .90 that the offending bus was defendant's. This estimate is subject to modification in light of additional evidence, of course; but the most accurate estimate one can make of the likelihood the bus was the defendant's, in the absence of other evidence, is .90. While many commentators in the Bayesian camp agree with the holding in *Smith,* they do so on grounds of policy considerations unrelated to doubts about the evidentiary value of base rates.[22]

[14] Kaye, *The Limits of the Preponderance of the Evidence Standard: Justifiably Naked Statistical Evidence and Multiple Causation,* 2 AM. B. FOUND. RES. J. 487, 488 (1982).

[15] *See, e.g.*, Smith v. Rapid Transit, 317 Mass. 469, 58 N.E.2d 754 (1945).

[16] *E.g.*, Turner v. U.S., 396 U.S. 398, 414-16, *reh'g denied,* 397 U.S. 958 (1970); Sindell v. Abbott Labs, 26 Cal. 3d 588, 607 P.2d 924, 163 Cal. Rptr. 132 (1980).

[17] 317 Mass. 469, 58 N.E.2d 754 (1945).

[18] *See* Koehler & Shaviro, *supra* note 11.

[19] Cohen, *Subjective Probability and the Paradox of the Gatecrasher,* 1981 ARIZ. ST. L.J. 627, 633. *See also* Brilmayer & Kornhauser, *Review: Quantitative Methods and Legal Decisions,* 46 U. CHI. L. REV. 116 (1978).

[20] Koehler & Shaviro, *supra* note 11, offer a cogent critique of the anti-Bayesian position.

[21] "From the perspective of verdict accuracy, it is unjustifiable to ignore, by reason of its unspecificity, the best available base rates." *Id.*

[22] For example, they fear that allowing a party to prevail based solely on "naked statistical evidence" may lead to strategic behavior, in which case-specific evidence is suppressed by the party favored by the base rate, and other "feedback effects" involving opportunistic responses to the knowledge that such evidence will be used. These concerns would generally not apply where base rates are offered in conjunction with more particularized evidence; hence most of these commentators argue that "directly relevant" base rates should be admitted where they are not "naked." But even "non-naked" base rates are sometimes excludable on policy grounds. For

The use in trials of "indirectly relevant" base rates has been more common. Base rates of this type may be used to show the weight that should be accorded a piece of forensic evidence. For example, where the perpetrator and the defendant are shown to have the same blood type, the prosecutor often presents statistics on the frequency of that blood type in a relevant population to prove that the match is unlikely to have occurred by chance.[23] Statistics on the percentage of the population possessing a given blood group are routinely admitted in evidence in most states.[24] Statistics also have been admitted in conjunction with forensic evidence showing a match between samples of hair,[25] glass and paint,[26] fibers,[27] particles,[28] and teeth marks.[29]

2. *Sources of Base Rate Statistics.* Because the value of associative evidence depends, in part, on the rarity of the characteristic or trace that links the defendant to the crime, forensic scientists have devoted much effort in recent years to studying the rarity of characteristics likely to be important in criminal identification. Efforts are being made in the United States and the United Kingdom to collect and store frequency data in a central location.[30] Base rate statistics literature is increasingly finding its way into criminal trials.[31]

The studies in this area are of two types. One type of study simply reports the relative frequency of various characteristics or traces in a sample drawn from some population. Most studies on the frequency of serological

example, it would be inconsistent with the constitutionally based presumption of innocence for a prosecutor to present base rate evidence showing that a high percentage of defendants in similar cases are convicted in order to show that a particular defendant is likely to be guilty.

[23] P. GIANNELLI & E. IMWINKELRIED, *supra* note 2, at 605-31; N. Miller, *supra* note 10.

[24] P. GIANNELLI & E. IMWINKELRIED, *supra* note 2, at 589-92, 605-38; Jonakait, *supra* note 6, at 639; Annotation, *Admissibility, Weight and Sufficiency of Blood Grouping Tests in Criminal Cases,* 2 A.L.R. 4th 500 (1980). The major exception, for a number of years, was New York, where in 1970 the state's highest appellate court found error in the admission of evidence that a defendant and perpetrator shared a blood type (Type A) found in 40% of the population. People v. Robinson, 27 N.Y.2d 864, 265 N.E.2d 543, 317 N.Y.S.2d 19 (1970); *accord* People v. Macedonio, 42 N.Y.2d 944, 366 N.E.2d 1355, 397 N.Y.S.2d 1002 (1977). The *Robinson* court found that this evidence was "of no probative value in the case against defendant in view of the large proportion of the general population having blood of this type" and expressed concern that jurors might give such evidence more weight than it deserves. 27 N.Y.2d at 865, 265 N.E.2d at 543, 317 N.Y.S.2d at 20. The court subsequently admitted evidence of a match on a set of blood group markers found in 1% of the population, however, arguing that "the relative rarity of the . . . type of blood relegates arguments as to remoteness to the realm of weight rather than admissibility." *In re* Abe A, 56 N.Y.2d 288, 299, 437 N.E.2d 265, 271, 452 N.Y.S.2d 6, 12 (1982). Then, in 1985, the court disavowed *Robinson,* citing near unanimous opposition to the holding by commentators, *e.g.*, MCCORMICK ON EVIDENCE § 205, at 619 (3d ed. 1984), and other courts, and recognizing that while proof of a match on a common characteristic has little value by itself, such evidence "may acquire great probative value when considered cumulatively." People v. Mountain, 66 N.Y.2d 197, 203, 486 N.E.2d 802, 805, 495 N.Y.S.2d 944, 947-48 (1985).

[25] *E.g.*, United States *ex rel.* DiGiacomo v. Franzen, 680 F.2d 516 (7th Cir. 1982).

[26] State v. Menard, 331 S.W.2d 521 (Mo. 1960).

[27] People v. Trujillo, 32 Cal. 2d 105, 194 P.2d 681 (1948).

[28] People v. Coolidge, 109 N.H. 403, 260 A.2d 547 (1969), *rev'd on other grounds,* 403 U.S. 443, *reh'g denied,* 404 U.S. 874 (1971).

[29] State v. Garrison, 120 Ariz. 255, 585 P.2d 563 (1978).

[30] Saferstein, *Criminalistics—A Look Back at the 1970s, A Look Ahead to the 1980s,* 24 J. FORENSIC SCI. 925 (1979).

[31] P. GIANNELLI & E. IMWINKELRIED, *supra* note 2, at 423-504, 605-31.

characteristics are of this type.[32] Studies have also been undertaken to determine the frequency of various types of paint,[33] glass,[34] fibers,[35] and soil;[36] the frequency of wear characteristics in men's footwear; and the frequency with which blood and semen stains[37] and glass and paint particles[38] are found on outer clothing and shoes.

A second type of study looks directly at the likelihood of a coincidental match between samples rather than the proportion of various characteristics in the population. Two Canadian forensic scientists, for example, conducted a study in which thousands of hairs from 100 unrelated individuals were compared under a microscope with respect to twenty-three different characteristics, such as color, pigment distribution, diameter, and scale count. In approximately one of every 4500 comparisons of hairs from different individuals, a match was found with respect to all twenty-three characteristics. Hence, the researchers reported that the chances of a coincidental match between two unrelated individuals on a microscopic comparison of scalp hairs is one in 4500.[39] Based on a subsequent study, the probability of a coincidental match for pubic hairs was estimated to be one in 800.[40] Data on the probability of coincidental match have also been collected on dental characteristics, DNA print patterns,[41] and even lipstick—the finding that there

[32] *E.g.*, Grunbaum, Selvin, Myhre & Pace, *Distribution of Gene Frequencies and Discrimination Probabilities of 22 Human Blood Genetic Systems in Four Racial Groups*, 25 J. Forensic Sci. 428 (1980); Steadman, *Blood Group Frequencies of Immigrant and Indigenous Populations for South East England*, 25 J. Forensic Sci. Soc'y 95 (1985).

[33] Ryland, Kipec & Somerville, *The Evidential Value of Automobile Paint, Part II: Frequency of Occurrence of Topcoat Color*, 26 J. Forensic Sci. 64 (1981).

[34] Fong, *Value of Glass as Evidence*, 18 J. Forensic Sci. 398 (1973).

[35] Home & Dudley, *A Summary of Data Obtained From a Collection of Fibres From Casework Materials*, 20 J. Forensic Sci. Soc'y 253 (1980).

[36] P. Giannelli & E. Imwinkelried, *supra* note 2, at 1080-86.

[37] Owen & Smalldon, *Blood and Semen Stains on Outer Clothing and Shoes Not Related to Crime: Report of a Survey Using Presumptive Tests*, 20 J. Forensic Sci. 391 (1975).

[38] Pearson, May & Dabbs, *Glass and Paint Fragments Found in Men's Outer Clothing-Report of a Survey*, 16 J. Forensic Sci. 283 (1971).

[39] Gaudette & Keeping, *An Attempt at Determining Probabilities in Human Scalp Hair Comparison*, 19 J. Forensic Sci. 599, 604 (1974). This study has been heavily criticized. *See, e.g.*, Barnett & Ogle, *Probabilities and Human Hair Comparison*, 27 J. Forensic Sci. 272 (1982); Note, *Splitting Hairs in Criminal Trials: Admissibility of Hair Comparison Probability Estimates*, 1984 Ariz. St. L.J. 521.

[40] Gaudette, *Probabilities and Human Pubic Hair Comparisons*, 21 J. Forensic Sci. 514, 517 (1976).

[41] DNA typing procedures produce "prints" consisting of a pattern of bands somewhat analogous to a supermarket bar code. *See generally* Thompson & Ford, *DNA Typing: Admissibility and Weight of the New Genetic Identification Tests*, 75 Va. L. Rev. 45 (1989). Individuals differ in the position of the bands on their print. To determine the likelihood of a coincidental match between DNA prints of two unrelated individuals, where the prtnts in question had fifteen distinct bands, Jeffreys, Wilson, and Thein made prints of twenty unrelated individuals, laid the prints side-by-side, and counted the number of instances in which a band in one print was matched by a band in the adjacent print. Jeffreys, Wilson & Thein, *Individual-specific "Fingerprints" of Human DNA*, 316 Nature 76 (1985). Overall, about 21% of the bands were matched by a band on an adjacent print. Accordingly, Jeffreys and colleagues concluded that there is about a 21% chance that a given band in a DNA print will be matched by a band in the print of an unrelated individual. To calculate the probability that two unrelated individuals will match on all fifteen bands, the researchers applied the product rule and concluded that the probability of a coincidental match on fifteen bands is approximately 0.21^{15} or one in thirty billion.

was a one-in-707 chance that samples of lipstick chosen at random would match was reportedly used as evidence in two criminal cases.[42]

3. *Drawing Conclusions from Base Rate Statistics: Potential Problems and Complexities.* Although base rate statistics are often highly relevant and informative, their probative value depends on a number of factors. To deal competently with base rate statistics, jurors must take these factors into account. However, several potential problems with base rate statistics may make these statistics misleading if jurors fail to understand their defects.

One problem is that base rates may be derived from inaccurate or uninformative data. For example, regional or geographic variations in the frequency of various types of fibers, paints, or soil types may render data based on samples in one area unrepresentative of frequencies in other areas. Kaye notes that for blood typing "the sampling is sufficiently extensive and variegated that the statistic should be reliable" while for other types of forensic evidence "scientific knowledge of the population parameter usually is . . . more sketchy."[43] Kaye favors admitting into evidence even these "sketchy" statistics on the grounds that they "can provide some clue as to the frequency of trace evidence in the population at large."[44] To draw appropriate conclusions from such statistics, however, jurors must appreciate the implications of sampling variability and sample bias. Research on the judgmental ability of untrained individuals raises some doubts about jurors' competence in this area.[45]

A second potential problem concerns computing the frequency of the joint occurrence of multiple characteristics. Where forensic evidence shows a match on several characteristics—for example, three distinct genetic markers in blood—forensic experts typically present statistics on the joint frequency of those characteristics—for example, the proportion of the population that possesses all three. These statistics are typically estimated from data on the frequency of the individual markers rather than from direct observation. Forensic scientists operate on the assumption that the genetic markers they use are independent of one another.[46] Accordingly, they compute the frequency of a combination of genetic markers by applying the product rule, which holds that the frequency of several independent events occurring simultaneously may be determined by simply multiplying the probability that each event will occur. If a match is found on three markers occurring in 5, 10,

[42] Barker & Clarke, *Examination of Small Quantities of Lipsticks,* 12 J. FORENSIC SCI. SOC'Y 449 (1972).

[43] Kaye, *supra* note 10, at 162. Suppose, for example, that the defendant is linked to the crime by the presence of red clay soil on his boots matching the soil at the scene of the crime. Statistics on the prevalence of red clay soil will be relevant only if they apply to the areas where the defendant might have been. The frequency of a given type of soil in a study in California may not be representative of the frequency of that soil type in Georgia.

[44] *Id.*

[45] R. NISBETT & L. ROSS, HUMAN INFERENCE: STRATEGIES AND SHORTCOMINGS IN JUDGMENT 77-88, 256-61 (1980).

[46] Independence presumes that possession of a given phenotype on one marker system is not associated with the possession of any particular phenotype on any other system.

and 20 percent of the population respectively, they typically report to the jury that the percentage of the population possessing all three markers is .05 x .10 x .20 = .001 or 0.1 percent.

The use of the product rule is well accepted for computing the frequency of protein markers in blood because there is extensive evidence that these markers are independent of one another.[47] Use of the product rule is inappropriate, however, where the characteristics are not independent. If the product rule is applied to events which are partially dependent, it may significantly underestimate the frequency of their joint occurrence. A number of courts have refused to allow computations based on the product rule unless the proponent can show that the characteristics being multiplied are independent,[48] but there are exceptions.[49] Hence, jurors sometimes must evaluate whether the use of the product rule to compute the statistics in the case was appropriate; where it is not, they must somehow take that problem into account.

In some areas the courts must decide whether to admit computations based on the product rule in the face of scientific uncertainty about the independence of the relevant characteristics. For example, considerable controversy has developed recently over the use of the product rule to compute the frequency of so-called DNA fingerprints when conclusive data have not appeared demonstrating that the genetic markers that make up the print are independent of one another.[50] While some courts have excluded the DNA statistics on this ground,[51] others have admitted the statistics, holding that any dispute over their accuracy should go to weight rather than admissibility.[52] Consequently, in a number of cases in which DNA evidence was presented, the issue of the independence of DNA markers has been thrown to the jury.[53]

[47] P. GIANNELLI & E. IMWINKELRIED, *supra* note 2, at 605.

[48] *E.g.*, People v. Collins, 68 Cal. 2d 319, 438 P.2d 33, 66 Cal. Rptr. 497 (1968).

[49] *E.g.*, State v. Garrison, 120 Ariz. 255, 585 P.2d 563 (1978) (product rule applied to determine the likelihood of matching bite marks in absence of demonstration of independence of matching features).

[50] In hearings on the admissibility of DNA statistics, some scientists have argued forcefully that the relevant markers ought to be independent, but others have questioned whether this opinion should be accepted in the absence of data to demonstrate its truth. Lander, *DNA Fingerprinting on Trial*, 339 NATURE 501, 503-04 (1989); Thompson & Ford, *Is DNA Fingerprinting Ready for the Courts?*, 125 NEW SCIENTIST 38, 43 (1990).

[51] *See, e.g.*, State v. Schwartz, 447 N.W.2d 422 (Minn. 1989) (ruling DNA test results by a commercial laboratory inadmissible based, in part, on the laboratory's failure to comply with a request for information about its data base that would have allowed the defense to assess the independence of the genetic markers); State v. Pennell, IN88-12-0051 (Del. Super. Ct. Sept. 20, 1989) (1989 WL 167430) (ruling statistics reported by the same lab inadmissible based, in part, on the failure of the laboratory to produce adequate documentation for its claim that the markers are independent).

[52] *E.g.*, People v. Wesley, 140 Misc. 2d 306, 533 N.Y.S.2d 643 (Albany County Ct. 1988). *See* Thompson & Ford, *supra* note 41, at 81-87 for a thorough discussion of this issue.

[53] The jurors' evaluation of the scientific evidence concerning independence may be crucial in such cases. In a hearing on the admissibility of DNA evidence in People v. Axell, CR23911 (Ventura Super. Ct. May 22, 1989), experts for the prosecution, who assumed independence and applied the product rule, testified that the frequency of defendant's DNA print was approximately one in 6 billion; experts called by the defense challenged the assumption of independence and testified that if

A third potential problem arises because the probative value of associative evidence (that is, evidence linking the defendant to the crime by showing a match) sometimes depends on how the defendant has been selected as a suspect. Where the defendant is selected for reasons unrelated to the likelihood of a match linking him to the crime, the frequency of the matching characteristic in the population from which the defendant was drawn is a reasonable estimate of the likelihood of a coincidental match. In other words, the frequency would provide the rough likelihood that the defendant would have the characteristic if innocent. If one person in 100 has the blood type on which the defendant and perpetrator match, for example, the probability is 1 percent that the defendant would happen by chance to have this blood type, if he is innocent. The frequency of the characteristic thus provides an index of the likelihood of a coincidental match. Where the defendant is selected for reasons that render him more or less likely than most people to have the matching characteristic, however, the frequency of the characteristic in the population does not reflect the likelihood of a coincidental match. If jurors fail to appreciate this fact, they may misjudge the likelihood of a misidentification and thereby over- or underestimate the value of the associative evidence.

As an illustration of this selection phenomenon, imagine a hypothetical murder case in which a long red hair (presumably from the killer) is found clenched in the fist of the victim. The police apprehend the defendant because he lives near the victim and has long red hair. Microscopic analysis reveals a match between hair samples taken from the defendant and the hair in the victim's hand with respect to thirteen distinguishable qualities such as color, length, and coarseness. A forensic expert reports a research study showing that the likelihood of a coincidental match between two hairs randomly drawn from different people is one in 4500.[54] What is the likelihood the defendant's hair would happen to match, as it does, if he is not the source of the hair in the victim's fist? Assuming one trusts the forensic report and the research, one is tempted to conclude the likelihood of a coincidental match is one in 4500, but this is demonstrably wrong. The defendant was selected, at least in part, because he has long red hair; thus, the likelihood that his hair would match the hair in the victim's fist, even though he is innocent, is undoubtedly far higher than the likelihood a randomly drawn individual would match. The figure of one in 4500 greatly underestimates the likelihood of a coincidental match in this case.

Perhaps the most blatant example of a selection effect occurred in the infamous case of *People v. Collins.*[55] A robbery was committed by a black man with a beard and a mustache and a blond woman with a ponytail, who both fled in a yellow convertible. Defendants were a couple fitting this description

the markers are not independent the frequency of the defendant's DNA print could be as high as one in 50. *See id.* testimony of expert Lawrence Mueller.

[54] For an actual case with similar facts, see State v. Carlson, 267 N.W.2d 170 (Minn. 1978).

[55] 68 Cal. 2d 319, 438 P.2d 33, 66 Cal. Rptr. 497 (1968).

apprehended in the vicinity of the robbery. To bolster a shaky eyewitness identification, the prosecutor called to the stand a college mathematics instructor and asked him to compute the frequency in the general population of a couple having various characteristics possessed by defendants, for example, a man with a beard, a man with a mustache, a blond woman, a woman with pony tail, and a yellow convertible. The prosecutor supplied "conservative" estimates of the frequency of these characteristics, and the mathematician, applying the product rule, multiplied the frequencies together to obtain a joint frequency of one in twelve million.

The California Supreme Court reversed the resulting conviction, concluding that it was error to admit the frequency estimate when that figure was not only likely to be overvalued by the jury but was computed on the unsupported assumption that the characteristics were independent.[56] Although the evidence in *Collins* certainly suffers from these problems, the major difficulty with the one in twelve million figure is that it purports to measure the probability of a coincidental match between the defendants and perpetrators but in fact does nothing of the sort. Assume that the mathematician was correct in computing the frequency of a couple with the stated characteristics to be one in twelve million. This figure might reflect the likelihood of a coincidental match if the defendants had been selected for reasons unrelated to the likelihood that they would possess the "matching" characteristics. However, it clearly does not reflect the likelihood of a coincidental match in the actual case, where defendants were selected precisely because they possessed the relevant characteristics. The likelihood of a match if these defendants are innocent is not one in twelve million, it is one in one; that is to say that it is certain.[57]

Courts and commentators have generally distinguished statistical testimony like that in *Collins* from testimony regarding the frequency of characteristics identified by forensic tests.[58] As the long red hair example illustrates, however, the same sort of selection effect that renders the *Collins* statistics problematic may also operate in cases involving forensic evidence, albeit in a more subtle manner.[59] As a result, base rate statistics in these cases

[56] *Id.* at 327, 438 P.2d at 38, 66 Cal. Rptr. at 502.

[57] If correct, the one in 12 million figure is not irrelevant; it suggests a low likelihood that another such couple would be found in the area and thus supports the conclusion that defendants are the guilty couple. In a city of several million people, however, the likelihood of finding two couples matching defendants' description might be reasonably high. In *Collins* the probability of a second couple in the Los Angeles area matching the characteristics of the perpetrators was computed to be .40. *Id.* at 333-35, 438 P.2d at 42-43, 66 Cal. Rptr. at 506-07.

[58] "[E]xpert testimony [in *Collins*] merely told the jury how to think—how to evaluate the fact that the Collins's were in the vicinity of the crime . . . . Such 'no evidence' cases do not dictate the outcome when meaningful statistical evidence permits a computation of the probability of a coincidental misidentification." Kaye, *supra* note 10, at 167.

[59] This problem is particularly likely to occur in cases in which forensic evidence shows a match on some observable characteristic (e.g., hair, paint, fibers). In such cases the suspect is often selected in a manner that renders him more likely than most people to "match" with regard to the relevant characteristic. On the other hand, where the "match" is on a characteristic that is not easily observed (e.g., blood type), it is less likely (though not inconceivable) that the characteristic played a role in

will not always reflect the probability of a coincidental misidentification. A key issue with regard to jury competence is whether jurors appreciate and take into account such phenomena.

4. *Presentation of Base Rate Statistics.* The difficulty jurors face in interpreting base rate statistics is compounded because these statistics may be presented in several different ways. Where forensic evidence shows a match, experts in most cases simply report the frequency of the matching characteristic or set of characteristics in a reference population, using percentages or incidence rates. It is common for experts to report, for example, that "type O blood is found in 44 percent of Caucasians" or that "among Hispanics, three persons in 100 possess both ABO type O and PGM-Type 2 enzyme markers." But other formulations are sometimes used. In *People v. Harbold,*[60] for example, serologist Mark Stolorow testified regarding the probability of a coincidental match between *two* individuals: " 'the chances of selecting any two people at random from the population and having them accidently [sic] have identical blood types in each one of these factors is less than one in 500, that is, what we call the probability of an accidental match is less than one in 500.' "[61] While it is tempting to assume that the frequency of a characteristic is equivalent to what Stolorow calls the probability of an accidental match, this is not the case. The probability of an accidental match actually equals the square of the frequency. If 10 percent of the population has blood type B, for example, the probability of selecting two people at random and finding they both have type B is .10 x .10 = .01, or one in 100. Thus, the impressive sounding conclusion that there is one chance in 500 of an accidental match is equivalent to the somewhat less impressive statement that the matching characteristics would be found in approximately one person in twenty-two.[62] Whether jurors appreciate this distinction between frequency and probability of accidental match is unclear.

To complicate matters, the probability of an accidental match on a genetic marker system is not necessarily equivalent to the probability of an accidental match on a particular marker. Consider, for example, the frequency of markers in the well-known ABO system. The probability that two randomly chosen individuals will share the same ABO type (not taking into account which type it is) is approximately 38 percent, while the probability they will both share a specific type ranges from 19 percent for type A to 0.16 percent for type AB.[63] Hence, it is crucial to know whether one is referring to an

the selection of the suspect and therefore less likely that the probability of a coincidental match diverges from the frequency of the matching characteristic.

[60] 124 Ill. App. 3d 363, 464 N.E.2d 734 (1984).

[61] *Id.* at 381, 464 N.E.2d at 748.

[62] It is possible that the underlying data Stolorow wished to report actually indicated a frequency of one in 500 for blood characteristics in question and that Stolorow mistakenly assumed that this frequency was equivalent to the probability of an accidental match. The appellate opinion leaves this point unclear.

[63] The frequency of ABO types and the probability of an "accidental match" with respect to each type (and any type) is shown in the following table:

accidental match on a system or on a specific marker within that system. But it may be difficult to tell from testimony such as Stolorow's[64] which of these terms is being reported. In addition, several different terms might be described in language that sounds similar. Suppose, for example, that the defendant and perpetrator share blood type AB. An expert might be quite correct in stating any of the following: (1) the probability of a match between two randomly chosen people in this system is 38 percent; (2) the probability of match between two randomly chosen people on this marker is 0.16 percent and (3) the probability a randomly chosen individual will have this marker is 4 percent. Whether jurors can and will appreciate the differences among these similar-sounding formulations is difficult to predict, but clearly this is an important issue underlying jury competence to deal with such data.

The difficulty jurors face in correctly interpreting base rate statistics becomes even greater when those statistics are reported in a misleading manner. Appellate opinions provide examples of erroneous and misleading statistical presentations. One error is to use base rate data to characterize the probability that *someone other than the defendant* was the source of an evidentiary sample. In *State v. Carlson,*[65] for example, forensic hair expert Barry Gaudette testified that there was "a 1-in-800 chance that the pubic hairs stuck to the victim were not Carlson's and a 1-in-4,500 chance that the head hairs found on the victim were not Carlson's."[66]

The problem with Gaudette's testimony is that it draws conclusions about the probability that the hairs "were not Carlson's." Obviously the hairs are either Carlson's or someone else's, so if Gaudette is correct in reporting one chance in 4500 the hairs were not Carlson's, it follows that there is a 4499 in 4500 chance they were Carlson's. Gaudette cannot properly testify to this effect, however, because conclusions about the likelihood that the hairs were Carlson's cannot be drawn from the forensic evidence alone. If one person in 4500 would have hair matching that found on the victim, thousands of people, besides Carlson, must have such hair. To determine the likelihood the hair

TABLE 1
FREQUENCY OF ABO TYPES

Type	Frequency in Population	Probability of "Accidental Match"
O	.44	.19
A	.42	.18
B	.10	.01
AB	.04	.0016
Overall (any type)		.3816

Because the frequency of type O is .44, the probability that two individuals drawn at random will both be type O is $.44^2 = .19$; that both will be type A is $.42^2 = .18$; that both will be type B is $.10^2 = .01$; and that both will be AB is $.04^2 = .0016$. Accordingly, the probability that two individuals will both have the *same* ABO type is $.19 + .18 + .01 + .0016 = .3816$. Selvin & Grunbaum, *Genetic Marker Determination in Evidence Bloodstains: The Effect of Classification Errors on Probability of Non-discrimination and Probability of Concordance*, 27 J. FORENSIC SCI. SOC'Y 57 (1987).

[64] *See supra* text accompanying note 60.

[65] 267 N.W.2d 170 (Minn. 1978).

[66] *Id.* at 173. *See* Gaudette & Keeping, *supra* note 39, at 605; Gaudette, *supra* note 40, at 517.

was Carlson's, we must consider whether the hair is more likely to have come from Carlson than from one of the thousands of other people with matching hair. One cannot make this determination, however, without evaluating the other evidence against Carlson, and therein lies the problem. Gaudette was in no position to evaluate the strength of the other evidence against Carlson and had no business doing so in any case. Hence, his opinion about the likelihood the hair was not Carlson's is not only unwarranted, but it also invades the province of the jury in a particularly insidious way, because the jurors are unlikely to realize that Gaudette's statistics rest, in part, on assumptions about the strength of evidence unrelated to the hair.

B. Error Rate Statistics

A second type of statistical formulation jurors may encounter in criminal trials concerns the rate of error in forensic tests.

1. *Sources of Error Rate Statistics.* The major source of error rate statistics is proficiency testing. A typical proficiency test is a blind trial in which forensic analysts are asked to classify specimens of known origin in order to check their accuracy. Studies of this type have provided considerable evidence that forensic testing is less than perfectly reliable. In 1974, the Law Enforcement Assistance Administration ("LEAA") of the Justice Department undertook a large-scale study of the proficiency of crime labs in the United States.[67] Between 235 and 240 laboratories took part in the blind trial, and the results, in the words of one commentator, were "shocking."[68] Over 20 percent of the labs inaccurately or incompletely identified samples of hair and paint, while over 30 percent of the labs inaccurately or incompletely identified glass and soil samples. Furthermore, less than 30 percent accurately or completely identified one sample of blood.

Error rates in blood typing are probably the best documented. Nationwide proficiency tests were conducted by the LEAA in 1975 and by the Forensic Sciences Foundation between 1978 and 1983.[69] Hundreds of blood samples of known type were sent to crime laboratories, which were asked to classify the samples while remaining "blind" to their type. The rate of classification errors varied among the different genetic marker systems used, ranging from 0.3 percent for the Adenosine deaminase system to over 6 percent for the familiar ABO system.[70] Because crime labs typically "type" blood on up to eight different systems, the likelihood of an error cumulates. Based on the proficiency test data, Selvin and Grunbaum concluded that

[67] J. PETERSON, E. FABRICANT & K. FIELD, CRIME LABORATORY PROFICIENCY TESTING RESEARCH PROGRAM—FINAL REPORT TO U.S. DEPT. OF JUSTICE (1978).

[68] Imwinkelried, *A New Era in the Evolution of Scientific Evidence: A Primer on Evaluating the Weight of Scientific Evidence*, 23 WM. & MARY L. REV. 261, 268 (1981).

[69] G. SENSABAUGH & D. NORTHEY, WHAT CAN BE LEARNED FROM THE PROFICIENCY TRIALS? AN ANALYSIS OF THE ELECTROPHORETIC TYPING RESULTS, 1975-83 PROCEEDINGS OF THE INTERNATIONAL SYMPOSIUM ON THE FORENSIC APPLICATIONS OF ELECTROPHORESIS 184 (1986). *See also* Selvin & Grunbaum, *supra* note 63, at 59.

[70] *See* Selvin & Grunbaum, *supra* note 63, at 59.

"blood group evidence employing [all] eight systems will be incorrect in some way in excess of 20 percent of the time."[71]

Blind trials have also been conducted recently to determine the proficiency of three commercial laboratories doing DNA typing of forensic samples. Asked to test approximately fifty unknown blood and semen samples, two of the labs had a "false-positive"; that is, they mistakenly declared a match in an instance where the pair of samples being compared actually came from different people.[72]

In addition to proficiency tests administered by outside agencies, many forensic laboratories engage in routine internal proficiency testing. These studies are potentially another source of data on error rates.

2. *Drawing Conclusions from Error Rate Statistics: Potential Problems and Complexities.* Like base rate statistics, error rate statistics are often highly relevant and informative, but must be interpreted with care because their probative value depends on a variety of factors. Jurors may be misled by such statistics if they fail to take these factors into account. However, the weight jurors should give to these factors is often unclear in a given instance. One important factor jurors should consider is whether aggregate data produced by proficiency testing of many labs accurately represent the rate of error in any particular lab. Large-scale proficiency tests such as those of the LEAA, for example, typically involve a number of laboratories, which are not individually identified. Consequently, it has been suggested that errors on proficiency tests result from inadequate training of a minority of analysts and tend to cluster in a few "bad" labs. As a result, aggregate data on error rates from proficiency tests greatly overstate the likelihood of an error by a competent analyst at a "good" lab.[73] A second consideration is whether error rates on proficiency tests reflect error rates in routine casework. In most proficiency tests, the laboratory personnel know they are being tested and may therefore be on their best behavior. Finally, jurors must consider whether error rates in the past predict the rate of errors in the future. One purpose of proficiency testing is to detect inadequacies in laboratory procedure that may contribute to error. Laboratories which have been "caught" making errors on proficiency tests sometimes change procedures in an effort to improve future performance.

[71] *Id.* at 61.

[72] One of the laboratories also had three false negatives, although these errors were initially covered up by the agency doing the testing. Ford & Thompson, *A Question of Identity: Some Reasonable Doubts About DNA "Fingerprints,"* THE SCIENCES, Jan./Feb. 1990, at 37, 41. The third lab had no false positives, but was unwilling to make a call on 14 of the samples and, in a follow-up study, twice failed to detect that mixed stains contained the DNA of two individuals. M. Graves & M. Kuo, DNA: A Blind Trial Study of Three Commercial Testing Laboratories (Feb. 1989) (paper presented at the meeting of the American Academy of Forensic Sciences, Las Vegas, Nev.).

[73] G. SENSABAUGH & D. NORTHEY, *supra* note 69. On the other hand, aggregate data may underestimate the rate of error at a "bad" lab. Hence, jurors must evaluate whether a particular lab is more or less error-prone than average to draw appropriate conclusions from aggregate error rate data.

A more subtle issue concerns the connection between the error rate of a test and the likelihood of a result that would falsely incriminate an innocent defendant. Not every error is of the sort that incriminates; therefore, the error rate of a test is not necessarily equivalent to the likelihood 'that an innocent person would be falsely incriminated. Suppose a bloodstain found at the scene of a crime is tested to see whether it matches that of a suspect, who is known to have type A blood. Assuming the stain is actually type O, the suspect will be falsely incriminated if the stain is misclassified as type A, but not if it is misclassified as type B or AB. The key issue, then, is not the overall error rate of the test but the rate at which types other than A are misclassified as type A. This error rate is sometimes called the false-positive rate for A. If errors are distributed randomly across the different blood types, the false-positive rate for a particular phenotype, such as type A, will be lower than the overall error rate for the test because only a subset of errors will be false-positives. If errors do not occur at random, however, the false-positive rate may be either higher or lower than the error rate. Suppose, for example, that there is an error rate of 6 percent in ABO typing, but that all of the errors occur when type O is misclassified as type A. In this instance, the false-positive rate for A would be higher than 6 percent while the false-positive rate for O, B, and AB would be zero. Although the connection between the error rate and the false-positive rate is not obvious in many instances, it is common for forensic scientists to report proficiency data in a form that allows inferences only about the overall error rate. The ability of jurors to draw appropriate conclusions from such data is open to question.

3. *Presentation of Statistics.* Although error rate statistics of this type are available in the published literature, they apparently are presented infrequently in criminal trials. A group of fifty forensic scientists surveyed by Miller[74] reported that they rarely presented data on error rates in court. Error rate data may be presented infrequently, in part, because attorneys are simply unfamiliar with it. Forensic scientists who are called to present the findings of forensic tests are unlikely to be examined extensively about error rates by the proponent of the evidence. While lawyers who cross-examine forensic experts are advised to probe extensively regarding the reliability of the procedure,[75] experts may be unwilling to phrase their estimates of error rates in statistical terms, or even to admit the possibility of error. In *State v. Spencer,*[76] for example, an expert responded to questions about the reliability of neutron activation analysis, a notoriously unreliable procedure,[77] by declaring "[t]here is no unreliability as far as we are concerned."[78] To challenge or even to detect such overstatements may require the attorney to

[74] N. Miller, *supra* note 10.

[75] E. IMWINKELRIED, THE METHODS OF ATTACKING SCIENTIFIC EVIDENCE (1982).

[76] 298 Minn. 456, 216 N.W.2d 131 (1974).

[77] George, *Statistical Problems Relating to Scientific Evidence*, in SCIENTIFIC AND EXPERT EVIDENCE 128 (E. Imwinkelried 2d ed. 1981).

[78] *Spencer*, 298 Minn. at 459, 216 N.W.2d at 134.

seek the assistance of another expert, whose services may be difficult to obtain or beyond the financial means of the defendant.[79]

Error rate statistics may also be used sparingly due to confusion about the meaning of errors on proficiency tests in relation to the reliability of a given procedure. Like data on the frequency of trace characteristics, error rate data are sketchy. Nevertheless, like frequency data, they provide some clue as to the likelihood of a wrong result. Whether jurors draw the appropriate conclusions from such data depends on their ability to appreciate the many subtle ways in which such statistics may be misleading. Whether lay individuals are capable of this task is an unexplored issue.

III

JURORS' USE OF STATISTICAL EVIDENCE: MAJOR CONCERNS AND RESEARCH STRATEGIES

Social scientists have recently begun studying whether lay individuals can draw appropriate conclusions from statistical evidence of the type presented in criminal trials. The standard research strategy is the jury simulation study, in which individuals read summaries of evidence and are asked to judge the guilt of a hypothetical criminal defendant. The nature of the evidence can be varied to determine, for example, how variations in the manner in which statistical evidence is presented affect people's judgments.

A major goal of the social scientists is to assess whether people use statistical evidence appropriately. To answer this question there must, of course, be some standard of appropriateness against which people's judgments can be compared. The benchmark used by researchers has been a set of mathematical models based on Bayes' theorem.[80] These models can specify how much one should revise one's estimate of a criminal suspect's probability of guilt after receiving forensic evidence accompanied by statistics.[81] Assuming a juror initially thinks that there is a 20 percent chance

[79] M. Saks & R. Van Duizend, *supra* note 1, at 89.

[80] For a general discussion of the use of Bayes' theorem to model legal judgments, see R. LEMPERT & S. SALTZBURG, A MODERN APPROACH TO EVIDENCE 148-53 (1st ed. 1977); Kaplan, *Decision Theory and the Factfinding Process*, 20 STAN. L. REV. 1065 (1968); Kaye, *What is Bayesianism? A Guide for the Perplexed*, 28 JURIMETRICS J. 161 (1988); Lempert, *Modeling Relevance*, 75 MICH. L. REV. 1021 (1977); Schum & Martin, *Formal and Empirical Research on Cascaded Inference in Jurisprudence*, 17 LAW & SOC'Y REV. 105 (1982).

[81] Where H and $\bar{H}$ designate the suspect's guilt and innocence respectively, and E designates evidence of a match between the suspect and perpetrator on some characteristic, Bayes' theorem states:

$$p(H/E) = p(H)p(E/H) / [p(H)p(E/H) + p(\bar{H})p(E/\bar{H})]$$

The term p(H) is read "the probability of H"; this term is called the prior probability and reflects the decisionmaker's initial estimate of the probability the suspect is guilty in light of everything that is known before receiving E. The term p(H/E) is read "the probability of H given E"; this term is called the posterior probability and indicates what the decisionmaker's revised estimate of probable guilt should be in light of everything known after receiving E. The formula indicates that the evidence of a match, E, should cause the decisionmaker to revise his opinion of the suspect's guilt to the extent p(E/H) differs from $p(E/\bar{H})$. If the suspect and perpetrator are certain to match if the suspect is guilty, p(E/H) = 1.00. If an innocent suspect is no more likely than anyone else to possess a particular suspect is guilty, for example, the models can tell him how much he should revise this estimate after receiving additional evidence that the suspect has genetic markers matching those found in the perpetrator's blood and that those markers occur in only 5 percent of the population.[82]

A major research strategy, then, is to determine whether people revise their judgments to the extent that Bayes' theorem dictates after receiving statistical evidence. Researchers are not particularly concerned with whether people's judgments correspond exactly to the predictions of Bayesian models. No one argues that jurors must be perfect intuitive Bayesians to be considered competent to deal with statistical data. Instead, the research has focused on three major concerns: first, whether people evaluate statistical evidence using inappropriate judgmental strategies that could lead to serious errors in estimating the value of the evidence, and therefore to dramatic divergence of human judgment from Bayesian norms; second, whether people are insensitive to important statistical variations in evidence and therefore fail to distinguish strong and weak evidence as effectively as the Bayesian models suggest they should; finally, whether people are insensitive to nonstatistical factors that affect the value of statistical evidence, such as partial redundancies between statistical evidence and other evidence in the case. By comparing actual judgments to those specified by Bayesian models, one can test sensitivity to such factors. In the sections that follow, each of these concerns will be discussed in some detail in light of the available empirical research and commentary.

A. Inappropriate Judgmental Strategies: Fallacious Interpretation of Statistical Evidence

1. *The Prosecutor's Fallacy.* One of the major concerns that has been raised about population proportions and statistics on the probability of a match is that jurors will mistakenly assume these statistics directly measure the probability of the defendant's innocence. A juror who hears that the defendant and perpetrator share a blood type found in 10 percent of the population, for example, may reason that there is only a 10 percent chance that the defendant would happen to have this blood type if innocent. The juror may then jump to the mistaken conclusion that there is therefore a 90 percent chance that the defendant is guilty. Thompson and Schumann,[83] who

the matching characteristic, $p(E/\bar{H})$ is equal to the frequency of the matching characteristic in the population from which the suspect was drawn.

[82] The prior probability of guilt, p(H) is equal to .20, and because the suspect must be either guilty or innocent, $p(\bar{H}) = .80$. Because the suspect is certain to have the perpetrator's genetic markers if he is the perpetrator, $p(E/H) = 1.00$; and because the suspect is no more likely than anyone else to have those genetic markers if he is not guilty, $p(E/\bar{H}) = .05$, the frequency of the markers in the population. These probabilites may be plugged into the Bayesian formula in note 81, *supra*, allowing one to solve for p(H/E), which in this case equals .83. In other words, learning that the suspect and perpetrator match on a characteristic found in 5% of the population should cause the decisionmaker to revise his estimate of likelihood of guilt from 10% to 83%.

[83] Thompson & Schumann, *supra* note 9.

call this mistake "the Prosecutor's Fallacy," explain the error by applying the underlying logic to a different problem:

> Suppose you are asked to judge the probability a man is a lawyer based on the fact he owns a briefcase. Let us assume all lawyers own a briefcase but only one person in ten in the general population owns a briefcase. Following the [fallacious] logic, you would jump to the conclusion that there is a 90 percent chance the man is a lawyer. But this conclusion is obviously wrong. We know that the number of nonlawyers is many times greater than the number of lawyers. Hence, lawyers are probably outnumbered by briefcase owners who are not lawyers (and a given briefcase owner is more likely to be a nonlawyer than a lawyer). To draw conclusions about the probability the man is a lawyer based on the fact he owns a briefcase, we must consider not just the incidence rate of briefcase ownership, but also the a priori likelihood of being a lawyer. Similarly, to draw conclusions about the probability a criminal suspect is guilty based on evidence of a "match," we must consider not just the percentage of people who would match but also the a priori likelihood that the defendant in question is guilty.[84]

The possibility that jurors might confuse population proportions with the probability of innocence was first raised by Laurence Tribe in his classic article, "Trial by Mathematics: Precision and Ritual in the Legal Process."[85] Discussing a murder case in which a partial palm print matching the defendant's is found on the murder weapon and a forensic expert testifies that such prints appear in no more than one case in a thousand, Tribe notes: "By itself, of course, the 'one-in-a-thousand' statistic is not a very meaningful one. It does not . . . measure the probability of the defendant's innocence—although many jurors would be hard-pressed to understand why not."[86] Tribe sees no problem with the admissibility of the forensic evidence of the match, but argues that the presentation of frequency data in connection with this evidence creates a serious danger of prejudice:

> To be sure, the finding of so relatively rare a print which matches the defendant's is an event of significant probative value, an event of which the jury should almost certainly be informed. Yet the *numerical index* of the print's rarity, as measured by the frequency of its random occurrence, may be more misleading than enlightening, and the jury should be informed of that frequency—if at all—only if it is also given a careful explanation that there might well be many other individuals with similar prints.[87]

Relying largely on Tribe's arguments, the Minnesota Supreme Court, in an interesting series of cases, has greatly limited the admissibility of statistics in connection with forensic evidence. No other appellate court has been as restrictive. The Minnesota opinions are worth examining in some detail because they raise a number of key issues about the competence of jurors to deal with statistics.

In *State v. Carlson*,[88] a rape and murder case in which hairs and semen were recovered from the victim, the court held that it was error (although nonprejudicial error) to admit statistical testimony on the probability of a match of characteristics. Specifically, the court found error in the admission of testimony by forensic hair expert Barry Gaudette that there was "a 1-in-800

[84] *Id.* at 170.

[85] Tribe, *supra* note 8.

[86] *Id.* at 1355.

[87] *Id.*

[88] 267 N.W.2d 170 (Minn. 1978).

chance that the pubic hairs stuck to the victim were not Carlson's and a 1-in-4,500 chance that the head hairs found on the victim were not Carlson's."[89]

Carlson argued that Gaudette's testimony "goes one step too far toward an ultimate conclusion of fact and therefore invades the province of the jury," and the court apparently agreed.[90] As the court saw it, however, the problem was not so much that Gaudette had used statistics improperly as that he had used statistics at all.

> Our concern over this evidence is not with the adequacy of its foundation, but rather with its potentially exaggerated impact on the trier of fact. Testimony expressing opinions ðr conclusions in terms of statistical probabilities can make the uncertain seem all but proven, and suggest, by quantification, satisfaction of the requirement that guilt be established "beyond a reasonable doubt." *See* Tribe, *Trial by Mathematics,* 84 HARV. L. REV. 1329 (1971).
>
> Diligent cross-examination may in some cases minimize statistical manipulation and confine the scope of probability testimony. We are not convinced, however, that such rebuttal would dispel the psychological impact of the suggestion of mathematical precision, and we share the concern for "the substantial unfairness to a defendant which may result from ill conceived techniques with which the trier of fact is not technically equipped to cope." People v. Collins, 68 Cal. 2d 332, 66 Cal. Rptr. 505, 438 P.2d 41. For these reasons we believe Gaudette's [statistical] testimony . . . was improperly received.[91]

Although Carlson left it unclear whether the court's objection is to all frequency data or only statistics on the probability of a match, a subsequent case, *State v. Boyd,*[92] reveals that the court was troubled by a broad range of statistical formulations.

In *Boyd,* a rape case, the prosecutor sought to show that the defendant had fathered the victim's child in order to prove he had achieved sexual penetration. Deciding a pretrial appeal of a trial court's decision to suppress the results of a paternity test, the court allowed evidence that a paternity test had failed to exclude the defendant as a possible father, but rejected accompanying statistical testimony on the percentage of men in the general population that the test would also exclude, as well as a statistical calculation of the "probability of paternity." The court again cited Tribe as support for its conclusion that

> there is a real danger that the jury will use the [statistical] evidence as a measure of the probability of the defendant's guilt or innocence, and that the evidence will thereby undermine the presumption of innocence, erode the values served by the reasonable doubt standard, and dehumanize our system of justice.[93]

The most recent case in this line is *State v. Joon Kyu Kim,*[94] another rape case, in which the prosecution appealed the trial court's decision to exclude statistics offered in connection with serological tests performed on the defendant and a semen sample extracted from the victim. The test results showed a match on a set of genetic markers that occur in only 3.6 percent of

[89] *Id.* at 173.
[90] *Id.* at 175.
[91] *Id.* at 176 (footnote omitted).
[92] 331 N.W.2d 480 (Minn. 1978).
[93] *Id.* at 483.
[94] 398 N.W.2d 544 (Minn. 1987).

the population. The frequency of the set of markers was computed by determining the frequency of each marker separately and then multiplying those frequencies together in accordance with the product rule. The court rejected the use of this frequency calculation on the grounds that it might be mistaken by the jury for the probability of Kim's innocence:

> [T]he expert called by the state . . . should not be permitted to express an opinion as to the probability that the semen is Kim's and should not be permitted to get around this by expressing the opinion in terms of the percentage of men in the general population with the same frequency of *combinations* of blood types.[95]

Retreating slightly from its previous rejection of all frequeney statistics, however, the court allowed testimony as to the percentage of men in the population who possess each of the individual matching genetic markers.[96] The court apparently believed these constituent probabilities were less likely to be prejudicial.

2. *Underutilization of Statistical Evidence.* In striking contrast to the concerns of Tribe and the Minnesota Supreme Court about the prejudicial potential of statistical evidence, other commentators have raised the opposite concern—that jurors will give statistics too little weight. Saks and Kidd criticise Tribe's analysis, calling it "a Swiss cheese of assumptions about human behavior—in this case human decision-making processes— which are asserted as true simply because they fall within the wide reach of the merely plausible, not because any evidence is adduced on their behalf."[97] Based on an extensive review of psychological studies on human judgment and decisionmaking, Saks and Kidd challenge arguments that statistics are inordinately persuasive and suggest that the reverse is true.[98]

A major psychological finding underlying Saks and Kidd's conclusion is the existence of the so-called base rate fallacy: the tendency for people, when judging the likelihood of an event, to ignore or underutilize statistical information on the base rate frequency of the event.[99] This tendency has been observed in a large number of studies.[100] When asked to judge whether a man described in a short vignette is a lawyer or an engineer, for example, people are nearly as likely to say he is a lawyer when told he was selected at random from a group consisting of 70 lawyers and 30 engineers as when told he was selected from a group consisting of 30 lawyers and 70 engineers.[101] The base rate data have relatively little impact on the judgment. "Only at the extremes of the distributions, where the group approaches 100 lawyers and 0

[95] *Id.* at 549.

[96] *Id.*

[97] Saks & Kidd, *supra* note 8, at 125.

[98] *Id.* at 149.

[99] For reviews, see 3 E. BORGIDA & N. BREKKE, THE BASE-RATE FALLACY IN ATTRIBUTION AND PREDICTION: NEW DIRECTIONS IN ATTRIBUTION RESEARCH (1981); Bar-Hillel, *The Base-Rate Fallacy in Probability Judgments*, 44 ACTA PSYCHOLOGICA 211 (1980).

[100] 3 E. BORGIDA & N. BREKKE, *supra* note 99; Bar-Hillel, *supra* note 99.

[101] Kahneman & Tversky, *On the Psychology of Prediction*, 80 PSYCHOLOGICAL REV. 237, 241-43 (1973).

engineers (or the converse) do the decision makers become sensitive to the information about group composition."[102]

The example of the base rate fallacy most relevant to jury competence is people's response to the well-known cab problem developed by psychologists Daniel Kahneman and Amos Tversky:[103]

> A cab was involved in a hit-and-run accident at night. Two cab companies, the Green and the Blue, operate in the city. You are given the following data:
>
> (a) 85 percent of the cabs in the city are Green and 15 percent are Blue.
>
> (b) a witness identified the cab as Blue . . . .
>
> [U]nder the same circumstances that existed on the night of the accident . . . the witness correctly identified each one of the two colors 80 percent of the time and failed 20 percent of the time.
>
> What is the probability that the cab involved in the accident was Blue rather than Green?[104]

This problem, and a number of similar problems, have been posed in dozens of psychological research studies.[105] As Saks and Kidd report, the typical probability response is 80 percent, although in actuality, the evidence given leads to a probability of 41 percent that the responsible cab was blue.[106] By judging the probability to be 80 percent, people are, in effect, ignoring the low base rate of blue cabs.

People's insensitivity to base rates in hypothetical problems of this type leads Saks and Kidd to suggest, in direct contradiction to Tribe and the Minnesota Supreme Court, that jurors are likely to pay little heed to base rates in actual legal proceedings. "[S]tatistical data need not be regarded as so overwhelming as some have supposed, and therefore they ought not to be considered prejudicial. The more realistic problem is presenting statistical evidence so that people will incorporate it into their decisions at all."[107]

It is important to notice, however, that the cab problem, as well as other problems revealing a "base rate fallacy," concerns the use of what this article has called "directly relevant" base rates, while the type of statistical evidence of concern to the Minnesota Supreme Court is "indirectly relevant" base rates. Thompson and Schumann have suggested that the two types of statistics "are likely to play a different role in the people's inferences" and therefore that the tendency to underutilize directly relevant base rates may

[102] Saks & Kidd, *supra* note 8, at 128.

[103] *See* Tversky & Kahneman, *Causal Schemata in Judgments Under Uncertainty*, in PROGRESS IN SOCIAL PSYCHOLOGY 49 (M. Fishbein ed. 1980); Tversky & Kahneman, *Evidential Impact of Baserates*, in JUDGMENT UNDER UNCERTAINTY: HEURISTICS AND BIASES (D. Dahneman, P. Slovic & A. Tversky eds. 1982) [hereinafter *Evidential Impact*].

[104] Tversky & Kahneman, *Evidential Impact*, *supra* note 103, at 156-57.

[105] *See* 3 E. BORGIDA & N. BREKKE, *supra* note 99; Bar-Hillel, *supra* note 99.

[106] Saks & Kidd, *supra* note 8, at 128. The correct solution to the problem may be obtained by applying Bayes' theorem. *See supra* notes 81, 82. The prior probability that the cab was blue, p(H), is .15 (and p(H) = .85) because 15% of the cabs in the city are blue. This prior probability must be revised in light of evidence, E, that the witness identified the cab as blue. Because the witness is accurate 80% of the time, p(E/H) - .80 and $p(E/\bar{H}) = .20$. Plugging these values into the Bayesian formula in note 81, *supra*, one can solve for p(H/E) and see that it is .41.

[107] Saks & Kidd, *supra* note 8, at 149.

not generalize to indirectly relevant base rates.[108] Thus, although Saks and Kidd marshal impressive evidence that statistics in general, and directly relevant base rates in particular, tend to be underutilized,[109] their analysis does not rule out the possibility that jurors may sometimes give too much weight to indirectly relevant base rates by falling victim to the Prosecutor's Fallacy of confusing the frequency of a matching characteristic with the probability of innocence.

3. *Jury Simulation Studies.* Additional light has been cast on the issue by recent studies of how simulated jurors react to statistical evidence when judging the guilt of hypothetical criminal defendants.[110] In a typical study, mock jurors are asked to estimate the probability that a hypothetical defendant is guilty based on a description of the evidence against him. The jurors are then asked to revise their initial estimate after receiving additional evidence indicating that the defendant and perpetrator share a characteristic, for example, a blood type, and specifying the statistical frequency of that characteristic. The weight jurors give to the evidence of the match is inferred from the extent to which they revise their estimates of probability of guilt after receiving it. A juror who initially thought the probability of guilt was 10 percent but revised his estimate to 90 percent after hearing about the match has given more weight to the evidence than a juror who revised the initial estimate from 10 percent to 15 percent. One advantage of assessing the weight mock jurors give to evidence in this manner is that it allows a direct comparison of human judgments to Bayesian predictions.[111] By using this approach, researchers can identify situations in which people give more or less weight to evidence than would be specified by Bayesian norms. It is also possible to detect instances in which people fall victim to the Prosecutor's Fallacy by determining whether they equate the frequency of a matching characteristic with the probability of innocence. In a case where the defendant and perpetrator matched on a characteristic found in 2 percent of the population, for example, a victim of the fallacy would always say the probability of guilt was 98 percent, regardless of the strength of the other evidence.

The studies generally find that some mock jurors make judgments consistent with the Prosecutor's Fallacy, although victims of this fallacy appear to be a small minority in most instances. For example, in the first study of this type, Thompson and Schuman asked 144 undergraduates to read a description of a hypothetical case involving the robbery of a liquor store by a man wearing a ski mask.[112] The police apprehended a suspect near the store who matched the store clerk's description of the robber's height, weight, and

[108] Thompson & Schumann, *supra* note 9, at 169.

[109] Saks & Kidd, *supra* note 8, at 132-45, 148-49.

[110] Faigman & Baglioni, *supra* note 9; Thompson & Schumann, *supra* note 9; J. Goodman, *supra* note 9; E. Schumann & W. Thompson, *supra* note 9.

[111] *See supra* notes 81-82.

[112] Thompson & Schumann, *supra* note 9, at 172-76.

clothing. In a trash can nearby, the police found the mask and the money. Subjects were asked, at this point, to give an initial estimate of the probability of the suspect's guilt. Then they read a summary of testimony by a forensic expert who reported that hair found in the ski mask matched the suspect's hair and that the frequency in the general population of hair that would match the suspect's was 2 percent. The subjects then made a final estimate of the suspect's probability of guilt. On average, about 13 percent of subjects judged the suspect's probability of guilt to be exactly 98 percent, which is the probability one would obtain by simply assuming that the frequency of the matching characteristic equals the probability of innocence. These subjects' comments during debriefing confirmed that they had fallen victim to the Prosecutor's Fallacy.[113] In other studies of this type, the percent of subjects making judgments consistent with the Prosecutor's Fallacy has been lower, ranging from about 1 percent to 8 percent.[114]

These studies also find evidence of a second fallacy, which Thompson and Schumann have labeled the "Defense Attorney's Fallacy."[115] This fallacy is the erroneous assumption that evidence of a match between the defendant and perpetrator on a rare characteristic is irrelevant to the defendant's likelihood of guilt.[116] For example, in a case where the defendant and perpetrator match on a characteristic found in 1 percent of the population, victims of this fallacy might reason that in a city of one million people there would be approximately 10,000 people with the relevant characteristic. They then might erroneously conclude that there is little, if any, probative value in the fact that the defendant and perpetrator both belong to such a large group. This reasoning is fallacious because the great majority of the 10,000 people with the relevant blood type are not suspects in the case and because the blood-test evidence drastically narrows the group of individuals who are or could be suspects without eliminating the very individual on whom suspicion has already focused. Victims of the Defense Attorney's Fallacy give no weight to evidence of a match on a rare characteristic. Consequently, in simulation studies their initial and final judgments of probability of guilt are identical. In the same study by Thompson and Schumann, described above, in which 13 percent of subjects made judgments consistent with the Prosecutor's Fallacy, another 12 percent made judgments consistent with the Defense Attorney's Fallacy.[117] Hence, one quarter of the subjects made judgments consistent with fallacious reasoning.

Susceptibility to the fallacious reasoning appeared to depend in part on how the statistical evidence was presented. When the frequency of the

[113] *Id.* at 173 n.5.

[114] In a second study reported by Thompson and Schumann, only 4% of subjects made judgments consistent with the prosecutor's fallacy. *Id.* at 177-81. Two similar studies reported by Goodman found rates of 1.6% and 8%. J. Goodman, *supra* note 9, at 8, 20.

[115] Thompson & Schumann, *supra* note 9, at 171.

[116] *Id.*

[117] *Id.* at 173.

matching characteristic was presented as a conditional probability,[118] the percentage of judgments consistent with the Prosecutor's Fallacy was higher (22 percent) and the percentage consistent with the Defense Attorney's Fallacy was lower (8 percent).[119] When the frequency was presented as a percentage and incidence rate,[120] the percentage of judgments consistent with the Prosecutor's Fallacy dropped to 4 percent but the percentage consistent with the Defense Attorney's Fallacy rose to 17 percent.[121]

Although these findings are suggestive, their practical significance is difficult to judge without more information. One limitation of the materials used in the study is that they included no arguments about the way statistical evidence should be used. In actual cases, jurors are likely to hear such arguments from attorneys or from other jurors during deliberation, so it is important to determine how people respond to them. Can people recognize the flaw in an argument for a fallacious position? Suppose they hear two fallacious arguments for contrary positions. How will they respond?

These questions were explored in a second study by Thompson and Schumann that was similar to the first[122] except that before subjects made final judgments of guilt, they received arguments.[123] One argument advocated the Prosecutor's Fallacy:

> The blood test evidence is highly relevant. The suspect has the same blood type as the attacker. This blood type is found in only 1% of the population, so there is only a 1% chance that the blood found at the scene came from someone other than the suspect. Since there is only a 1% chance that someone else committed the crime, there is a 99% chance the suspect is guilty.[124]

The other argument advocated the Defense Attorney's Fallacy:

> The evidence about blood types has very little relevance for this case. Only 1% of the population has the "rare" blood type, but in a city . . . [l]ike the one where the crime occurred with a population of 200,000 this blood type would be found in approximately 2000 people. Therefore the evidence merely shows that the suspect is one of 2000 people in the city who might have committed the crime. A one-in-2000 chance of guilt (based on the blood test evidence) has little relevance for proving *this* suspect guilty.[125]

Half the subjects read the prosecutor's argument first followed by the defendant's, while the other half read the arguments in reverse order. After reading each set of arguments, subjects were asked to indicate whether they thought either was correct and to judge the suspect's probability of guilt.

Most subjects found at least one of the fallacious arguments convincing. Twenty-nine percent thought the argument for the Prosecutor's Fallacy was

[118] The expert stated there was "only a two percent chance [that] the defendant's would be indistinguishable from that of the perpetrator if he were innocent . . . ." *Id.*

[119] *Id.* at 174.

[120] The expert stated that 2 percent of people have hair that would be indistinguishable and that in a city of 1,000,000 people there would be approximately 20,000 such individuals. *Id.* at 173.

[121] *Id.* at 174.

[122] The case involved evidence of a match between the suspect and perpetrator on a blood type found in 1% of the population. *Id.* at 177-81.

[123] *Id.* at 177.

[124] *Id.*

[125] *Id.* at 178.

correct.[126] Sixty-eight percent thought the argument for the Defense Attorney's Fallacy was correct.[127] Only 22 percent correctly concluded that both arguments were incorrect.[128]

Not surprisingly, after hearing the arguments a much higher percentage of subjects made judgments consistent with fallacious reasoning than in the earlier experiment. However, the Defense Attorney's Fallacy seemed to dominate. Over 50 percent of the subjects made judgments of probable guilt consistent with the Defense Attorney's Fallacy, giving no weight to evidence of a match on a characteristic found in 1 percent of the population,[129] but only 4 percent made judgments consistent with the Prosecutor's Fallacy.[130]

Overall, these findings suggest that people have difficulty detecting fallacious arguments—especially the argument favoring the Defense Attorney's Fallacy—and that these arguments can lead significant numbers of people to make judgments consistent with fallacious reasoning. In other words, it is easy to talk people into using inappropriate judgmental strategies to evaluate "indirectly relevant" base rate evidence presented in conjunction with forensic evidence.

With the exception of individuals who fall victim to the Prosecutor's Fallacy, however, most subjects in these studies appeared to give less weight to evidence of a match than Bayes' theorem says they should. A consistent finding, observed in six experiments,[131] is that subjects, on average, revise judgments of probability of guilt upward by a smaller amount than that required by Bayesian norms. In the first study by Thompson and Schumann, for example, subjects' initial judgments of probability of guilt averaged about 25 percent.[132] According to Bayes' theorem, after learning of a match on a characteristic found in 2 percent of the population, subjects' final judgments should have been about 93 percent.[133] However, subjects' actual judgments averaged only 63 percent,[134] indicating that, on average, they gave the evidence of the match less weight than they should have. This result is typical of findings in other studies.[135]

Researchers in this area sometimes warn that their findings should be viewed as preliminary.[136] Many of the studies are rather rudimentary simulations of trials in which subjects read summaries of evidence rather than see actual testimony. In addition, the subjects make individual judgments instead of deliberating as a group. Subjects are often asked to judge

[126] *Id.*
[127] *Id.*
[128] *Id.* at 178-79.
[129] *Id.*
[130] *Id.*
[131] Faigman & Baglioni, *supra* note 9; Thompson & Schumann, *supra* note 9, at 176, 180, experiments 1, 2; J. Goodman, *supra* note 9, studies 1, 2; Schumann & Thompson, *supra* note 9.
[132] Thompson & Schumann, *supra* note 9, at 174.
[133] *Id.* at 175.
[134] *Id.*
[135] Faigman & Baglioni, *supra* note 9; J. Goodman, *supra* note 9, at 30.
[136] *See, e.g.*, Thompson & Schumann, *supra* note 9, at 183.

probability of guilt rather than to decide whether to convict or acquit, hence some concerns have been expressed about whether findings of such studies "[go] beyond the articulation of numbers and actually [influence] the sorts of decisions juries are called upon to make."[137]

Recently, however, researchers have begun conducting more realistic studies. For example, Schumann and Thompson[138] recently examined the effects of fallacious statistical arguments in the context of a highly realistic simulated trial. Subjects in the role of jurors viewed a four-hour videotape of a simulated trial based on transcripts of an actual California murder case. Although some of the testimony was abbreviated or replaced with stipulations, the simulated trial included virtually all of the evidence presented in the case on which it was based. The attorneys in the simulated trial were, in fact, experienced criminal lawyers, and the judge was a real judge who gave legally appropriate instructions.

The case involved a robbery and murder in which the perpetrator was injured, leaving blood at the scene. Although a considerable amount of circumstantial evidence was presented, the case against the defendant was weak except for a key piece of forensic evidence—genetic markers in his blood matched those of the blood at the scene, and the relevant markers are found in only 2 percent of the population.

Five different versions of the trial were shown to 116 simulated jurors. Some jurors heard the prosecutor make an argument advocating the Prosecutor's Fallacy, while others heard the prosecutor state only the frequency of the genetic markers. Within each of those two groups, half of the subjects heard the defense attorney make an argument for the Defense Attorney's Fallacy and half did not. Thus one group heard competing fallacious arguments, a second group heard no fallacious arguments, a third group heard just the argument for the Prosecutor's Fallacy, and a fourth heard just the argument for the Defense Attorney's Fallacy. The fifth group was a control condition in which evidence of the matching blood types and the accompanying statistical evidence were not presented. The arguments lasted less than one minute in the context of a ten- to fifteen-minute closing argument in a four-hour trial.

After watching the trial, the simulated jurors individually indicated whether they would vote guilty or not guilty and estimated the probability that the defendant actually committed the crime. Then they deliberated in groups of six for up to an hour before again indicating their choice of verdict and estimate of probability of guilt.

Before deliberation, half of the jurors voted guilty although no significant differences were found among the five conditions.[139] In other words, the evidence of the matching blood markers appeared to have no effect, regardless of how it was argued. Effects of the arguments emerged after

[137] *Id.* at 183.

[138] Schumann & Thompson, *supra* note 9.

[139] *Id.* at 5.

deliberation, however. The conviction rate and estimates of probability of guilt dropped significantly in all conditions except the condition where jurors heard only the argument for the Prosecutor's Fallacy.[140] In that condition, the conviction rate increased to 70 percent and average estimates of probability of guilt increased to 85 percent, while in the other conditions, the conviction rate ranged from 17 percent to 30 percent and average estimates of probability of guilt ranged from 51 to 59 percent.[141] The short argument for the Prosecutor's Fallacy thus had a powerful effect on jurors' ultimate verdicts, but only where it was not countered by the argument for the Defense Attorney's Fallacy. Jurors who heard arguments for both fallacies, for the Defense Attorney's Fallacy only, or for neither fallacy produced conviction rates and estimates of probability of guilt that did not differ significantly from one another and that were only slightly higher than those in the control condition.[142]

These findings suggest that an argument for the Prosecutor's Fallacy can be influential, even in the context of a highly realistic trial, if not countered by the defense. Overall, however, these findings offer little support for those who are concerned that jurors will give too much weight to statistical evidence. Although the simulated jurors gave some weight to the evidence of matching blood types—estimates of probability of guilt were significantly higher where that evidence was presented than in the control group—the difference between the control group and the conditions where the blood typing evidence was presented were not as great as a Bayesian analysis indicates they should have been. As in the previous studies, mock jurors appear generally to undervalue rather than overvalue this evidence. Indeed, subjects appeared to come closest to giving the evidence its correct value only in response to the uncountered argument for the Prosecutor's Fallacy. This finding raises an intriguing issue. How should the legal system respond if it turns out that the only way to induce jurors to give the proper weight to such evidence is to trick them with fallacious statistical arguments?

B. Sensitivity to Variations in Statistical Evidence

A second major concern that has been raised about the use of statistical evidence is that jurors may be insensitive to variations in important statistical values.[143] Jurors' sensitivity to statistical variation is an important issue because, under some circumstances, variations that may seem relatively minor in the frequency of a matching characteristic or the false-positive rate of the

[140] *Id.* at 5-6.

[141] *Id.*

[142] *Id.* at 5-10. If we take judgments of probability of guilt in the control condition as a prior probability and revise them in light of the forensic evidence in the Bayesian manner, we find the final judgments of probability of guilt should, on average, be above .90. Subjects' actual judgments were not that high in any condition, although those in the Prosecutor's Fallacy-only condition were close.

[143] Thompson, Britton & Schumann, *supra* note 9, at 2.

test showing the match can have a large effect on the probative value of the evidence.[144]

Suppose, for example, that a juror initially estimates the probability of the defendant's guilt to be only 10 percent, but then receives new, independent evidence indicating that the defendant and perpetrator have the same blood type. If the incidence rate of the blood type and the false-positive rate of the test showing the match are both 1 percent, then, according to the Bayesian model, the juror should revise the estimate of probability of guilt upward to about .85 percent. If the incidence rate and false-positive rate are both 5 percent, however, the Bayesian model indicates the juror's revised estimate of probability of guilt should be only .53. Although the difference between 1 percent and 5 percent may appear small, it has a dramatic impact on the probative value of the match between the defendant and perpetrator.[145]

To evaluate the results of a forensic test linking the defendant to the crime in cases where the test is less than perfectly reliable, jurors must evaluate two factors: the possibility that the test result showing a match is a false-positive, and the possibility that the match, if correct, is merely coincidental. Experts on human judgment have suggested that when faced with problems such as this involving two sources of uncertainty, people often proceed in a stepwise fashion, following what has become known as the "best guess" strategy.[146] To reduce the complexity of the judgment, people make their best guess as to whether the evidence is reliable and, if they think it is probably reliable, they proceed to evaluate the evidence as if it were perfectly reliable. They then discount their certainty about their conclusions to take into account their uncertainty about the reliability of the evidence. They often fail to discount this evidence adequately, however. The result is that judgments based on less than fully reliable evidence are often unduly extreme because of the failure to discount adequately for unreliable evidence.[147] If this process influences jurors' evaluations of forensic matching evidence, it could cause jurors to be insensitive to variations in the reliability of the evidence.

The sensitivity to variations in the statistics accompanying forensic matching evidence was examined directly in a series of studies reported by Thompson, Britton, and Schumann.[148] In the standard experimental paradigm, mock jurors were asked to evaluate evidence of a match between the suspect and the perpetrator on a rare blood type. The rarity of the blood type and the false-positive rate of the forensic test were both experimentally varied to be either 1 or 5 percent. In the first study, undergraduate mock

[144] The way a juror should respond to forensic evidence involving frequency statistics and false-positive rates may be specified by a mathematical model based on Bayes' theorem. *Id.* at Appendix. Thompson, Britton, and Schumann derived this model and used it as a benchmark against which to compare actual judgments based on such data.

[145] *Id.* at 7.

[146] *See, e.g.*, Gettys, Kelly & Peterson, *The Best-Guess Hypothesis in Multistage Inference*, 10 ORG. BEHAV. & HUM. PERFORMANCE 364 (1973); Slovic, Fischhoff & Lichtenstein, *Behavioral Disorder Theory*, 28 ANN. REV. PSYCHOLOGY 1 (1977).

[147] Gettys, Kelly & Peterson, *supra* note 146.

[148] Thompson, Britton & Schumann, *supra* note 9.

jurors were given several hypothetical cases at once in which the frequency and false-positive rate varied. These subjects were able to evaluate accurately the relative strength of the evidence. In other words, they could tell that the evidence was strongest where the frequency and false-positive rate were low, weakest where the frequency and false-positive rate were high, and of intermediate value where one factor was high and the other low. In the second study, however, subjects were given only one piece of forensic evidence to evaluate and the frequency and false-positive rate statistics were experimentally varied among different groups of subjects.[149]

Although the statistical variation should have made a large difference in the value of the evidence, the different groups of subjects did not differ significantly in the value they assigned to it.[150] The group that received the strongest evidence (low frequency, low false-positive rate) did not give the evidence more weight than the group that received the weakest evidence (high frequency, high false-positive rate). It thus appears that people can rank in order several pieces of evidence according to relative strength but have difficulty evaluating the absolute strength of any single piece of evidence.[151] Actual trials are, of course, more analogous to the second experiment than the first, because jurors are called upon to evaluate absolute strength of forensic matching evidence rather than the relative strength of several pieces of evidence.

In a third experiment, Thompson, Britton, and Schumann tested mock jurors sensitivity to these statistical variations in a more realistic study which included group deliberation.[152] Again they found that mock jurors did not differentiate weak from strong evidence. The hypothetical case was devised in such a way that Bayesian predictions of probability of guilt were 64 percent in the strong evidence condition, where forensic tests had a low false-positive rate and found a match on a rare characteristic, but only 22 percent in the weak evidence condition, where forensic tests had a higher false-positive rate and found a match on a more common characteristic.[153] The conviction rate of subjects in the two conditions did not significantly differ, however, and was actually a bit higher in the weak evidence condition (84 percent) than in the strong evidence condition (78 percent).[154] Following deliberation, subjects in both conditions estimated the probability of the defendant's guilt. Among subjects in the strong evidence condition the average estimate was 66 percent, which is very close to Bayesian norms. Among subjects in the weak evidence condition, however, the average estimate was 65 percent, which is much higher than the Bayesian prediction of 22 percent, indicating that

[149] *Id.* at 7-10.

[150] *Id.* at 10-19.

[151] *Id.* at 11.

[152] *Id.* at 12-13.

[153] *Id.* at 15-20.

[154] *Id.* at 17.

subjects' failure to differentiate weak and strong statistical evidence led them to overestimate the value of weak evidence in this instance.[155]

More research is needed in this area to test the generality of these provocative but rather preliminary findings. These studies suggest, however, that under some circumstances jurors may seriously overestimate the value of statistical evidence, as Tribe and the Minnesota Supreme Court feared. The reason for this problem, however, is not jurors' tendency to confuse the frequency of a matching characteristic with the probability of innocence, but their tendency to give equal weight to statistical evidence that varies widely in its probative value.

Recommendations for dealing with this problem may be a bit premature. Until the scope of the problem is better understood, attempted solutions might easily miss the mark. If these preliminary findings are borne out by further research, however, one possible solution would be for the courts to be especially cautious about admitting forensic evidence of questionable reliability. Forensic matching evidence that is relatively weak, by virtue of having a high false-positive rate, might be particularly likely to be overvalued.

C. Dealing With Partially Redundant Evidence

A third problem concerns subjects' reactions to partially redundant evidence, that is, evidence that partly overlaps and recapitulates facts they have already taken into account.[156] Forensic evidence showing a match between the suspect and perpetrator is partially redundant in cases where the suspect was selected in a manner that renders him more likely than most people to match the perpetrator on a certain characteristic. In the *Carlson* case[157] discussed earlier, for example, the defendant's hair matched, in a number of separate qualities, a hair taken from the perpetrator. Assuming the defendant was arrested in part because his hair matched the perpetrator's with regard to color and length, the evidence of a match on all qualities is partly redundant with what is already known.

The inferential complexities created by partial redundancy among multiple items of evidence have been discussed by philosophers of inductive logic,[158] legal evidence scholars,[159] and Bayesian theorists.[160] A variety of terms have been used to describe partially redundant evidence. The terms "cumulative" and "corroborative" are preferred by most legal scholars, but those terms will be avoided here because, as Schum notes, the precise

[155] *Id.* at 17-19, 25.

[156] *See supra* notes 55-59 and accompanying text.

[157] State v. Carlson, 267 N.W.2d 170 (Minn. 1978).

[158] J. VENN, THE PRINCIPLES OF INDUCTIVE LOGIC (2d ed. 1905); S. TOULMIN, THE USES OF ARGUMENT (1964).

[159] J. WIGMORE, THE SCIENCE OF JUDICIAL PROOF (1937); Lempert, *Modeling Relevance*, 75 MICH. L. REV. 1021 (1977).

[160] Most notably, D. Schum, On Factors which Influence the Redundancy of Cumulative and Corroborative Testimonial Evidence (1979) (Technical Report #79-02). *See also* Schum & Martin, *supra* note 80.

meaning varies so much among different scholars that their use promotes more confusion than clarity.[161] A precise account of the ways in which evidence may be partially redundant requires an appreciation of the catenated or "cascaded" nature of inductive inference.[162] Accordingly, the best accounts of partially redundant evidence are provided by Wigmore, who uses complex evidentiary diagrams to show connections among various pieces of evidence,[163] and Schum, who uses formal mathematical models of cascaded inference derived from Bayes' theorem.[164] Lempert also provides a clear, though less complete, account of partial redundancy using Bayesian models.[165]

Although considerable attention has been paid to formal descriptions of partial redundancy in evidence, relatively little is known about people's ability to apprehend and deal appropriately with this evidentiary subtlety when evaluating the evidence in criminal trials. The only empirical study that sheds light on this question was reported by Schum and Martin.[166] In this complex and sophisticated study, subjects evaluated evidence in hypothetical criminal cases in three different but formally equivalent ways. In some instances, subjects evaluated the evidence in the entire case "holistically;" in other instances the evidence was either partially or totally "decomposed," and subjects made separate evaluations of its constituent elements. Subjects were trained to evaluate the evidence by estimating likelihood ratios and conditional probabilities and they evaluated the evidence in these terms. The major finding of relevance here is that subjects' evaluations were more sensitive to partial redundancies among items of evidence when the evidence was decomposed than when evaluations were "holistic."[167] This finding raises the possibility that people may be inadequately sensitive to partial redundancy when, as in actual criminal trials, they are evaluating evidence that is not explicitly "decomposed" for them.

Some additional light has been cast on the problem by a study reported by Thompson, Meeker, and Britton.[168] Undergraduate subjects were asked to read written descriptions of evidence in a series of criminal cases. For each case, the description provided an account of the nature of the crime and the manner in which the suspect was identified. It then described two hypothetical pieces of forensic evidence that might be offered against the suspect. Subjects were asked to judge which of the two pieces of evidence would constitute stronger evidence of the suspect's guilt. Each piece of evidence revealed a match between the suspect and the perpetrator in a different but equally rare characteristic. One piece of evidence was partially

[161] Schum & Martin, *supra* note 80, at 117.

[162] *Id.* at 116-18.

[163] J. WIGMORE, *supra* note 159, at 154.

[164] Schum, *supra* note 160. *See also* Schum & Martin, *supra* note 80.

[165] Lempert, *supra* note 80. *See also* R. LEMPERT & S. SALTZBURG, *supra* note 80, at 148-53.

[166] Schum & Martin, *supra* note 80.

[167] *Id.*

[168] Thompson, Meeker & Britton, *supra* note 9.

redundant because the suspect had been selected in a way that rendered him unusually likely to have that matching characteristic even if innocent. The other piece of evidence was not redundant because the suspect was no more likely than anyone else to have it if he was innocent. For example, one case involved the burglary of a drug store by a black perpetrator who was injured, leaving blood at the scene. The police identified a suspect based, in part, on the fact that he was black. Subjects were then asked which of two pieces of evidence would be stronger: (1) evidence that the suspect and perpetrator both have sickle-cell characteristic, a trait found in about one person in 100 in the United States, but most commonly found among blacks and rarely found in other races, or (2) evidence that the suspect and perpetrator both have a hypothetical genetic characteristic (factor Q), which is found in one person in 100 in the United States but is evenly distributed among the races. The match on factor Q is clearly stronger evidence against the suspect because he was identified based in part on a characteristic (his race) that renders him more likely than the general population to have sickle-cell trait if innocent.

The goal of the study was simply to test whether people realize that the partially redundant piece of evidence deserved less weight than the nonredundant evidence. In general, it appears that they do not.[169] In most of the hypothetical cases, about a third of subjects thought the partially redundant evidence was stronger, about a third thought the two pieces of evidence were equally strong, and a third thought (correctly) that the nonredundant evidence was stronger. This distribution of responses is what would be expected if people could not detect any difference between the partially redundant and nonredundant evidence and simply responded at random. People's ability to appreciate the distinction appears to improve following discussion of the issue with others. Even after discussing the issue for up to twenty minutes with others, however, approximately half of the subjects still chose the incorrect option when asked which piece of evidence was stronger.[170]

Although these findings are preliminary, they raise serious concerns about the ability of jurors to detect partial redundancies in forensic evidence and to take those into account. As a result, jurors may overvalue forensic matching evidence in cases in which it is partially redundant with other evidence.

IV

CONCLUSION

Empirical research on peoples' evaluation of statistical evidence, although preliminary and full of gaps, is beginning to define the strengths and weaknesses of lay statistical reasoning in ways that should prove helpful to the legal system. This research casts some much-needed light on a number of issues that have divided commentators and troubled appellate courts.

[169] *Id.*

[170] *Id.*

However, it appears that the broad question posed by this article has no single or simple answer.

The research should allay fears that jurors will overvalue statistical evidence by mistakenly equating the frequency of matching characteristics with the probability of innocence. Although some jurors do fall victim to this type of fallacious reasoning, they are a small minority in all studies and the prevailing tendency is toward undervaluing rather than overvaluing such evidence. Arguments for the Prosecutor's Fallacy can have a powerful influence on judgments of guilt, but can be countered effectively by opposing arguments.

Jurors may sometimes overvalue forensic evidence used to link a defendant to a crime, but this potential problem arises not from the Prosecutor's Fallacy but from people's failure to take into account the unreliability and partial redundancy of forensic evidence. Where the value of forensic matching evidence is significantly undermined by the high false-positive rate of a forensic test, and where the defendant was selected in a manner that renders him more likely to possess the matching characteristics than the general population, there appears to be a significant danger that the forensic evidence will be overvalued. Until these judgmental tendencies are better understood, courts would be well advised to use caution when considering the admissibility of statistics in connection with such evidence.

Exhibit 13

Laurence H. Tribe, Trial by Mathematics: Precision and Ritual in the Legal Process

84 Harv. L. Rev. 1329

Harvard Law Review
April, 1971

Laurence H. Tribe[a1]

TRIAL BY MATHEMATICS: PRECISION AND RITUAL IN THE LEGAL PROCESS

Professor Tribe considers the accuracy, appropriateness, and possible dangers of utilizing mathematical methods in the legal process, first in the actual conduct of civil and criminal trials, and then in designing procedures for the trial system as a whole. He concludes that the utility of mathematical methods for these purposes has been greatly exaggerated. Even if mathematical techniques could significantly enhance the accuracy of the trial process, Professor Tribe also shows that their inherent conflict with other important values would be too great to allow their general use.

THE system of legal proof that replaced trial by battle in Continental Europe during the Middle Ages reflected a starkly numerical jurisprudence. The law typically specified how many uncontradicted witnesses were required to establish various categories of propositions, and defined precisely how many witnesses of a particular class or gender were needed to cancel the testimony of a single witness of a more elevated order.[1] So it was that medieval law, nurtured by the abstractions of scholasticism, sought in mathematical precision an escape from the perils of irrational and subjective judgment.

In a more pragmatic era, it should come as no surprise that the search for objectivity in adjudication has taken another tack. Yesterday's practice of numerology has given way to today's theory of probability, currently the *sine qua non* of rational analysis. Without indulging in the dubious speculation that contemporary ***1330** probabilistic methods will one day seem as quaint as their more mystical predecessors, one can at least observe that the resort to mathematical techniques as alternatives to more intuitive tools in the trial process has ancient roots. Nor is it entirely accidental that those roots seem oddly twisted when examined outside their native soil. For, although the mathematical or pseudo-mathematical devices which a society embraces to rationalize its systems for adjudication may be quite comprehensible to a student of that society's customs and culture, those devices may nonetheless operate to distort--and, in some instances, to destroy--important values which that society means to express or to pursue through the conduct of legal trials. This article discusses the respects in which this is the case--and, in so doing, suggests a framework of analysis for the assessment of the potentialities and dangers of current and proposed uses of mathematical methods in the trial process.

In speaking of mathematical methods "in the trial process," I am referring to two related but nonetheless separable topics: not only to the use of mathematical tools in the actual conduct of a particular trial, but also to the use of such tools in the design of the trial system as a whole. The first topic encompasses such questions as the propriety of allowing the parties in a lawsuit to employ explicitly statistical evidence or overtly probabilistic arguments for various purposes,[2] and the wisdom of permitting or encouraging the trier to resolve the conflicting claims of a lawsuit with the assistance of mathematical methods.[3] The second topic, in contrast, centers on the desirability of employing such methods in establishing the procedural and evidentiary rules according to which lawsuits generally should be conducted. Both topics, of course, share a common core: both involve the wisdom of using mathematical tools to facilitate the making of choices among available courses of action with respect to the trial process. In this sense, both topics form part of the larger subject of when and how mathematical methods ought to ***1331** be employed in decisionmaking. And this subject, in turn, is part of the still more inclusive topic of when it is desirable to make decisions in a calculating, deliberate way, with the aid of precise and rigorous techniques of analysis. To the extent that this article sheds any light on those larger matters, I will of course be gratified. I will not, however, attempt to deal directly with them here, and will instead confine myself to the narrower inquiries outlined above.

Two further introductory remarks are in order. First, my subject is the use of mathematics as a tool for decisionmaking rather than simply as a mode of thought, as an instrument rather than as a language. Conceivably, the very enterprise of describing some phenomena in precise mathematical terms, and particularly the enterprise of quantifying them, might be shown to entail some significant costs in addition to its obvious benefits. Perhaps it is in some sense “dehumanizing” to talk in highly abstract or quantitative terms about some subjects,[4] but this is another issue not to be treated here.

Second, although my central concern is the wisdom of using mathematical methods for certain decisionmaking purposes even when those methods are rationally employed, I will also examine what must be regarded as clearly irrational uses of those methods. Thus, some might charge that, by relying on such misuses in any overall assessment, I have confused the avoidable costs of using a tool badly with the inherent costs of using it well. It is rather like the claim that statistics can lie. One may always respond that this claim is false while conceding that the devil can quote Scripture to his own purposes. In a sense, this is obviously the case. But in another sense, it is only a half-truth, for the costs of abusing a technique must be reckoned among the costs of using it at all to the extent that the latter creates risks of the former. To be more precise, in at least some contexts, permitting *any* use of certain mathematical methods entails a sufficiently high risk of misuse, or a risk of misuse sufficiently costly to avoid, that it would be irrational not to take such misuse into account when deciding whether to permit the methods to be employed at all.

Finally, a word about objectives. This analysis has been undertaken partly because I suspect that the lure of objectivity and precision may prove increasingly hard to resist for lawyers concerned with reliable, or simply successful, adjudication; partly because a critique of mathematical efforts to enhance the reliability and impartiality of legal trials may yield helpful insights into what such trials are and ought to be; and partly because such a ***1332** critique may ultimately contribute to an appreciation of how rigor and quantification, once their real costs and limits are better understood, might actually prove useful in processes of decisionmaking. Most fundamentally, though, I write in reaction to a growing and bewildering literature of praise for mathematical precision in the trial process,[5] a literature that has tended to catalogue or to assume the virtues of mathematical approaches quite as uncritically as earlier writers[6] tended to deny their relevance.

I. FACTFINDING WITH MATHEMATICAL PROBABILITIES

A. Mysteries of Expertise

The infamous trial in 1899 of Alfred Dreyfus, Captain in the French General Staff, furnishes one of the earliest reported instances of proof by mathematical probabilities. In attempting to establish that the author of a certain document that allegedly fell into German hands was none other than Captain Dreyfus, the prosecution called several witnesses who theorized that Dreyfus must have written the document in question by tracing the word *intérêt* from a letter written by his brother, constructing a chain of several of these traced words in a row, and then writing over this chain as a model when preparing the document--in order to give it the appearance of a forgery and thereby to protect himself should the document later be traced to him. [7] To identify the writing in the document as that of Dreyfus, the prosecution's witnesses reported a number of close matches between the lengths of certain words and letters in the document and the lengths of certain words and letters in correspondence taken from Dreyfus' home. Obscure lexicographical and graphological "coincidences" within the document itself were said by the witnesses to indicate ***1333** the high probability of its disguised character and of its use to convey coded information. [8] To establish the validity of the hypothesis that the document had been traced over the handwriting of Dreyfus' brother, the prosecution's witnesses computed the "amazing" frequency with which certain letters in the document appeared over the same letters of the word chain constructed by repeating *intérêt* a number of times, once a variety of complex adjustments had been made. [9]

The very opacity of these demonstrations protected them to some degree from effective spontaneous criticism, but the "mathematics" on which they were based was in fact utter nonsense. As the panel of experts appointed several years later to review the evidence in the Dreyfus case easily showed, [10] there was nothing statistically remarkable about the existence of close matches in some word lengths between the disputed document and Dreyfus' correspondence, given the many word pairs from which the prosecution was free to choose those that displayed such similarities. [11] Moreover, the supposed coincidences within the document itself reflected no significant deviation from what one would expect in normal French prose. Finally, the frequency with which various letters in the document could be "localized" over the letters of *intérêt* was likewise statistically insignificant.

Armand Charpentier, a prominent student of the Dreyfus affair, reports that counsel for Dreyfus and the Government Commissioner alike declared that they had understood not a word of the witness' mathematical demonstrations. [12] Charpentier adds that, although the judges who convicted Dreyfus were in all likelihood equally mystified, they nonetheless "allowed themselves to ***1334** be impressed by the scientific phraseology of the system." [13] It would be difficult to verify that proposition in the particular case, but the general point it makes is a crucial one: the very mystery that surrounds mathematical arguments--the relative obscurity that makes them at once impenetrable by the layman and impressive to him--creates a continuing risk that he will give such arguments a credence they may not deserve and a weight they cannot logically claim.

The California Supreme Court recently perceived this danger when it warned that "[m]athematics, a veritable sorcerer in our computerized society, while assisting the trier of fact in the search for truth, must not [be allowed to] cast a spell over him." [14] The court ruled improper a prosecutor's misconceived attempt to link an accused interracial couple with a robbery by using probability theory. The victim of the robbery, an elderly woman, had testified that she saw her assailant, a young woman with blond hair, run from the scene. One of the victim's neighbors had testified that he saw a Caucasian woman, with her hair in a dark blond ponytail, run from the scene of the crime and enter a yellow automobile driven by a male Negro wearing a mustache and beard. Several days later, officers arrested a couple that seemed to match these descriptions. [15] At the week-long trial of this couple, the victim was unable to identify either defendant, and her neighbor's trial identification of the male defendant was effectively impeached. [16] Moreover, the defense introduced evidence that the female defendant had worn light-colored clothing on the day of the robbery, although both witnesses testified that the girl they observed had worn dark clothing. Finally, both defendants took the stand to deny any participation in the crime, providing an alibi that was at least consistent with the testimony of another defense witness.

In an effort to bolster the identification of the defendants as the perpetrators of the crime, the prosecutor called a college mathematics instructor to establish that, if the robbery was indeed committed by a Caucasian woman with a blond ponytail accompanied ***1335** by a Negro with a beard and mustache and driving a yellow car, there was an overwhelming probability that the accused couple were guilty because they matched this detailed description.

The witness first testified to the “product rule” of probability theory, according to which the probability of the joint occurrence of a number of mutually independent events equals the product of the individual probabilities of each of the events.[17] Without presenting any supporting statistical evidence, the prosecutor had the witness assume specific probability factors for each of the six characteristics allegedly shared by the defendants and the guilty couple.[18] Applying the product rule to the assumed factors, the prosecutor concluded that there was but one chance in twelve million that any couple chosen at random would possess the characteristics in question, and asked the jury to infer that there was therefore but one chance in twelve million of the defendants' innocence.

The jury convicted but the California Supreme Court reversed, holding the mathematical testimony and the prosecutor's associated argument inadmissible on four separate grounds. First, the record was devoid of any empirical evidence to support the individual probabilities assumed by the prosecutor.[19]

Second, even if the assumed probabilities were themselves correct, their multiplication under the product rule presupposed the independence of the factors they measured--a presupposition for which no proof was presented, and which was plainly false.[20] If two or more events tend to occur together, the chances of their separate occurrence obviously cannot be multiplied to yield the chance of their joint occurrence.[21] For example, if every tenth ***1336** man is black and bearded, and if every fourth man wears a mustache, it may nonetheless be true that most bearded black men wear mustaches, so that nearly one man in ten--not one in forty-- will be a black man with a beard *and* a mustache.

Third, even if the product rule could properly be applied to conclude that there was but one chance in twelve million that a randomly chosen couple would possess the six features in question, there would remain a substantial possibility that the guilty couple did not in fact possess all of those characteristics--either because the prosecution's witnesses were mistaken or lying, or because the guilty couple was somehow disguised. “Traditionally,” the court reasoned, “the jury weighs such risks in evaluating the credibility and probative value of trial testimony,”[22] but-- finding itself unable to quantify these possibilities of error or falsification--the jury would be forced to exclude such risks from any effort to assign a number to the probability of guilt or innocence and would be tempted to accord disproportionate weight to the prosecution's computations.

Fourth, and entirely apart from the first three objections, the prosecutor erroneously equated the probability that a randomly chosen couple would possess the incriminating characteristics, with the probability that any given couple possessing those characteristics would be innocent. After all, if the suspect population contained, for example, twenty-four million couples, and if there were a probability of one in twelve million that a couple chosen at random from the suspect population would possess the six characteristics in question, then one could well expect to find two such couples in the suspect population, and there would be a probability of approximately one in two--not one in twelve million--that any given couple possessing the six characteristics would be innocent.[23] The court quite reasonably thought that few ***1337** defense attorneys, and fewer jurors, could be expected to comprehend these basic flaws in the prosecution's analysis.[24] Under the circumstances, the court concluded, this “trial by mathematics” so distorted the jury's role and so disadvantaged defense counsel as to constitute a miscarriage of justice.[25]

But the California Supreme Court discerned "no inherent incompatability between the disciplines of law and mathematics and intend[ed] no general disapproval ... of the latter as an auxiliary in the fact-finding processes of the former." [26] Thus expressed, the court's position seems reasonable enough. Any highly specialized category of knowledge or technique of analysis is likely to share in some degree the divergence between impressiveness and understandability that characterizes mathematical proof; surely, adjudication should not for that reason be deprived ***1338** of the benefits of all expertise. On the contrary, the drawing of unwarranted inferences from expert testimony has long been viewed as rectifiable by cross-examination, coupled with the opportunity to rebut. Particularly if these devices are linked to judicial power to give cautionary jury instructions and to exclude evidence altogether on a case-by-case basis if prejudicial impact is found to outweigh probative force, and if these techniques are then supplemented by a requirement of advance notice of intent to use a particular item of technical proof, and by some provision for publicly financed expert assistance to the indigent accused confronted with an expert adversary, [27] there might seem to be no valid remaining objection to probabilistic proof.

But can such proof simply be equated with expert evidence generally, or does it in fact pose problems of a more pervasive and fundamental character? A consideration of that question requires the more careful development of just what "mathematical proof" should be taken to mean, and what major forms it can assume.

B. Illustrative Cases: Occurrence; Identity; Intention

In an examination of the role of mathematical methods in the trial itself, whether used by one or more of the parties in the presentation of proof or employed by the trier in reaching a decision, we may set aside at the outset those situations in which the very issues at stake in a litigation are in some sense mathematical and hence require the explicit trial use of mathematical techniques--when, for example, the governing substantive law makes a controversy turn on such questions as percentage of market control, [28] expected lifetime earnings, [29] likelihood of widespread public confusion, [30] or the randomness of a jury selection process. [31] My concern is with cases in which mathematical methods are turned to the task of deciding what occurred on a particular, unique occasion, as opposed to cases in which the very ***1339** task defined by the applicable law is that of measuring the statistical characteristics or likely effects of some process or the statistical features of some population of people or events.

With this initial qualification in mind, it is possible--and will occasionally prove helpful--to separate mathematical proof into three distinct but partially overlapping categories: (1) those in which such proof is directed to the *occurrence* or nonoccurrence of the event, act, or type of conduct on which the litigation is premised; (2) those in which such proof is directed to the *identity* of the individual responsible for a certain act or set of acts; and (3) those in which such proof is directed to *intention* or to some other mental element of responsibility, such as knowledge or provocation. In dealing with the utility of mathematical proof in the trial process, I will later show how such a tripartite division can be useful. It is sufficient to say at this stage that the significance, appropriateness, and dangers of mathematical proof may depend dramatically on whether such proof is meant to bear upon occurrence, identity, or frame of mind. [32] Several examples should suffice to illustrate the contents of each of these categories.

1. Occurrence.--Consider first the cases in which the existence of the legally significant occurrence or act is itself in question. A barrel falls from the defendant's window onto the plaintiff's head. The question is whether some negligent act or omission by defendant caused the fall. Proof is available to support a finding that, in over sixty percent of all such barrel-falling incidents, a negligent act or omission was the cause. Should such proof be allowed and, if so, to what effect? [33]

A man is found in possession of heroin. The question is whether he is guilty of concealing an illegally imported narcotic drug. Evidence exists to support the finding that ninety-eight ***1340** percent of all heroin in the United States is illegally imported. What role, if any, may that fact play at the defendant's trial?[34]

A man is charged with overtime parking in a one-hour zone. The question is whether his car had remained in the parking space beyond the time limit. To prove that it had not been moved, the government calls an officer to testify that he recorded the positions of the tire air-valves on one side of the car. Both before and after a period in excess of one hour, the front-wheel valve was pointing at one o'clock; the rear-wheel valve, at eight o'clock. The driver's defense is that he had driven away during the period in question and just happened to return to the same parking place with his tires in approximately the same position. The probability of such a fortunate accident is somewhere between one in twelve and one in one hundred forty-four. [35] Should proof of that fact be allowed and, if so, to what end?[36]

2. Identity.--Consider next the cases in which the identity of the responsible agent is in doubt. Plaintiff is negligently run down by a blue bus. The question is whether the bus belonged to the defendant. Plaintiff is prepared to prove that defendant ***1341** operates four-fifths of all the blue buses in town. What effect, if any, should such proof be given?[37]

A policeman is seen assaulting someone at an undetermined time between 7 p.m. and midnight. The question is whether the defendant, whose beat includes the place of the assault, was the particular policeman who committed the crime. It can be shown that the defendant's beat brings him to the place of the assault four times during the relevant five-hour period each night, and that other policemen are there only once during the same period. In what way, if at all, may this evidence be used?[38]

A man is found shot to death in the apartment occupied by his mistress. The question is whether she shot him. Evidence is available to the effect that, in ninety-five percent of all known cases in which a man was killed in his mistress' apartment, the mistress was the killer. How, if at all, may such evidence be used?[39]

A civil rights worker is beaten savagely by a completely bald ***1342** man with a wooden left leg, wearing a black patch over his right eye and bearing a six-inch scar under his left, who flees from the scene of the crime in a chartreuse Thunderbird with two dented fenders. A man having these six characteristics is charged with criminal battery. The question is whether the defendant is in fact the assailant. Evidence is available to show that less than one person in twenty has any of these six characteristics, and that the six are statistically independent, so that less than one person in sixty-four million shares all six of them. In what ways, if at all, may that calculation be employed?[40]

3. Intention.--Consider finally the cases in which the issue is one of intent, knowledge, or some other “mental” element of responsibility. A recently insured building burns down. The insured admits causing the fire but insists that it was an accident. ***1343** On the question of intent to commit arson, what use, if any, may be made of evidence tending to show that less than one such fire out of twenty is in fact accidentally caused?[41]

As in an earlier example, [42] a man is found possessing heroin. This time the heroin is stipulated at trial to have been illegally imported. In his prosecution for concealing the heroin with knowledge that it had been illegally imported, what effect may be given to proof that ninety-eight percent of all heroin in the United States is in fact illegally imported?[43]

A doctor sued for malpractice is accused of having dispensed a drug without adequate warning, knowing of its tendency to cause blindness in pregnant women. Should he be allowed to introduce evidence that ninety-eight percent of all doctors are unaware of the side-effect in question?

4. An Overview.--The reader will surely note that this collection of cases might have been subdivided along a variety of different axes. Some of the cases are civil, others criminal. Some involve imputations of moral fault; others do not. Some rest upon statistical calculations that might readily be made; others, on figures that are at best difficult to obtain and at worst entirely inaccessible. Some entail the use of probabilistic evidence to establish liability; others, to negate it. In some, the probabilities refer to a party's own involvement in a category of events; in others, they refer to the proportion of similar events in which a certain critical feature is present, or in which the responsible party has a certain important characteristic. In some of these cases, the mathematics seems best suited to assisting the judge in his allocation of burdens of production or persuasion; in others, its most natural role seems to be as evidence for the finder of fact.

My aim in classifying the cases in terms of occurrence, identity, and intention is not to imply that these other ways of carving up the topic have less significance, but merely to sketch one possible map of the territory I mean to cover--using a set of boundaries that are intuitively suggestive and that will prove helpful from time to time as the discussion unfolds. [44]

Courts confronted with problems of the several sorts enumerated in the three preceding sub-sections have reacted to them on an almost totally ad hoc basis, occasionally upholding an attempt ***1344** at probabilistic proof, [45] but more commonly ruling the particular attempt improper. [46] A perhaps understandable pre-occupation with the novelties and factual nuances of the particular cases has marked the opinions in this field, to the virtual exclusion of any broader analysis of what mathematics can or cannot achieve at trial--and at what price. As the number and variety of cases continue to mount, the difficulty of dealing intelligently with them in the absence of any coherent theory is becoming increasingly apparent. Believing that a more general analysis than the cases provide is therefore called for, I begin by examining--and ultimately rejecting--the several arguments most commonly advanced against mathematical proof. I then undertake an assessment of what I regard as the real costs of such proof, and reach several tentative conclusions about the balance of costs and benefits.

C. The Traditional Objections

The cases sketched in the preceding section differ in many respects, but all of them share three central features which have at times been thought to preclude any meaningful application of mathematical techniques. The first of those is that, in all of these cases, concepts of probability based upon the relative frequency of various events must be applied, if at all, not to the statistical prediction of a possible *future* event but to a determination of the occurrence or characteristics of an alleged *past* event. At first glance, probability concepts might appear to have no application in deciding precisely what did or did not happen on a specific prior occasion: either it did or it didn't--period.

The New York Court of Appeals elevated that intuition into a rule of law when it rejected probabilistic testimony to show that a forgery had been done on the defendant's typewriter. [47] ***1345** The court distinguished the judicially accepted use of life expectancy tables on the ground that such use arises “from necessity when the fact to be proved is the probability of the happening of a future event. It would not be allowed,” the court continued, “if the fact to be established were whether *A* had in fact died, to prove by the Carlisle Table he should still be alive.” [48] Thus, the court reasoned, probabilistic testimony (as to the rarity of the coincidence between peculiarities in the defendant's typewriter and peculiarities in the forged document) should be disallowed since the “fact to be established … was not the probability of a future event, but whether an occurrence asserted by the people to have happened had actually taken place.” [49] The court's result was defensible on far narrower grounds,[50] but this reasoning is not. It is not the future character of an event that induces us to give weight to probabilistic evidence, but the lack of other, more convincing, evidence--an absence more common in, but ***1346** certainly not limited to, future occurrences.[51] Indeed, "sense-perception itself" might be viewed as "a form of prediction for action purposes,"[52] and "propositions about past facts … [can be regarded as] 'predictions,' on existing information, as to what the 'truth' will turn out to be when and if more knowledge is available."[53] Insofar as the relevance of probability concepts is concerned, then, there is simply no inherent distinction between future and past events. That all of the cases sketched in the preceding section would apply such concepts to a determination about a past occurrence therefore gives rise to no objection of substance.

However, a second similarity among the cases put above is less easily dismissed: in all of them, making use of the mathematical information available first requires transforming it from evidence about the *generality* of cases to evidence about the *particular* case before us. Some might suggest that no such transformation is possible, and that no translation can be made from probability as a measure of objective frequency in the generality of cases to probability as a measure of subjective belief in the particular instance. That suggestion would be incorrect.[54] In the bus case, to take a typical example, we start with the objective fact that four-fifths of the blue buses are operated by the defendant. That datum can obviously point to a correct conclusion in the particular case, for it suggests that, in the absence of other information, in some sense there is a "four-fifths certainty" that the defendant's bus hit this plaintiff. To be sure, the complete "absence of other information" is rare,[55] but the mathematical datum nonetheless provides a useful sort of knowledge--in part to guide the judge's allocation of the burden of producing believable evidence,[56] and in part to convey to the factfinder a relatively precise sense of the probative force of the background information that is available.

But does it really mean anything at all to be "four-fifths ***1347** certain" in a particular case? Unlike many such questions, this last, fortunately, has an answer--one first formulated rigorously by Leonard Savage in a seminal 1950 work.[57] Professor Savage, employing elegantly few assumptions, developed a "personalistic" or "subjective" theory of probability based on the notion that it makes sense to ask someone what he would do if offered a reward for guessing correctly whether any proposition, designated *X,* is true or false. If he guesses that *X* is true under these circumstances, we say that *for him* the subjective probability of *X,* written P(X), exceeds fifty percent. Symbolically, P(X) > .5. If he would be equally satisfied guessing either way, then we say that, for him, P(X) = .5.

As Professor Savage demonstrated, this basic concept can readily be expanded into a complete scale of probabilities ranging from 0 to 1, with P(X) = 0 representing a subjective belief that *X* is impossible and P(X) = 1 representing a subjective belief that *X* is certain.[58] Thus, one could take a sequence of boxes each containing a well-shuffled deck of one hundred cards, some marked "True" and others marked "False." The first box, B_0, would contain no cards marked "True" and 100 marked "False"; the second box, B_1, would contain 1 card marked "True" and 99 marked "False;" the next box, B_2, would contain 2 cards marked "True" and 98 marked "False"; and so on, until the last box, B_{100}, would contain 100 cards marked "True" and no cards marked "False." To say that a person is "four-fifths" or "eighty percent" certain of *X* is simply to say that he would be as willing to bet that the proposition *X* is true as he would be willing to bet (with the same stakes) that a card chosen at random from box B_{80} will turn out to be marked "True." In these circumstances, the person would say that, for him, P(X) = .8.

In the context of the bus case, if a person knew only that eighty percent of all blue buses are operated by the defendant,[59] and if he had to bet one way or the other, he should be as willing to bet that the bus involved *in this case* was defendant's as he would be willing to bet that a card chosen at random from B_{80} ***1348** would be marked "True."[60] This is what it means to say that, for him, the subjective probability of the defendant's liability [61] --pending the receipt of further information-- equals .8.

Different people, of course, would typically assign different subjective probabilities to the same propositions--but that is as it must be, unless the propositions in question are unusually simple. And at least in the law we do not find startling the notion that reasonable men with differing life experiences and differing assumptions will assess the same evidence differently.

The interesting thing about subjective probabilities as defined by Savage is that, once a few entirely plausible postulates are accepted, [62] these probabilities obey the usual rules that the schoolboy associates with such simple operations as flipping fair coins or drawing cards from a well-shuffled deck; hence the translation from objective frequencies to subjective probabilities called for by all of the cases we have considered can indeed be made. [63] ***1349** Again, there are no insuperable obstacles to the application of mathematical techniques.

Once the translation to subjective probabilities is completed, we encounter the third similarity among the cases. In very few of them, if any, can the mathematical evidence, *taken alone and in the setting of a completed lawsuit,* establish the proposition to which it is directed with sufficient probative force to prevail. To return once again to the blue bus litigation, [64] even assuming a standard of proof under which the plaintiff need only establish his case “by a preponderance of the evidence” in order to succeed, the plaintiff does not discharge that burden by showing simply that four-fifths, or indeed ninety-nine percent, of all blue buses belong to the defendant. [65] For, unless there is a satisfactory explanation for the plaintiff's singular failure to do more than present this sort of general statistical evidence, we might well rationally arrive, once the trial is over, at *a subjective probability of less than .5* that defendant's bus was really involved in the specific case. [66] And in any event, absent satisfactory explanation, there are compelling reasons of policy to *treat* the subjective probability as less than .5--or simply as insufficient to support a verdict for plaintiff. To give less force to the plaintiff's evidentiary omission would eliminate any incentive for plaintiffs to do more than establish the background statistics. The upshot would be a regime in which the company owning four-fifths of ***1350** the blue buses, however careful, would have to pay for *five*-fifths of all unexplained blue bus accidents--a result as inefficient as it is unfair. [67]

A fortiori, when the governing standard of proof is more stringent still, the mathematics taken alone would typically fall short of satisfying it. As the California Supreme Court put it in the *Collins* case, “no mathematical equation can prove beyond a reasonable doubt (1) that the guilty [party] *in fact* possessed the characteristics described by the People's witnesses, or even (2) that only *one* [party] possessing those characteristics could be found in the [relevant] area.” [68]

But the fact that mathematical evidence *taken alone* can rarely, if ever, establish the crucial proposition with sufficient certitude to meet the applicable standard of proof does not imply that such evidence--*when properly combined with other, more conventional, evidence in the same case*--cannot supply a useful link in the process of proof. Few categories of evidence indeed could ever be ruled admissible if each category had to stand on its own, unaided by the process of cumulating information that characterizes the way any rational person uses evidence to reach conclusions. The real issue is whether there is any acceptable way of combining mathematical with non-mathematical evidence. If there is, mathematical evidence can indeed assume the role traditionally played by other forms of proof. The difficulty, as we shall see, lies precisely in this direction--in the discovery of an acceptable integration of mathematics into the trial process. I now turn to a consideration of the only plausible mode of integration yet proposed.

D. A Possible Solution

In deciding a disputed proposition, a rational factfinder probably begins with some initial, a priori estimate of the likelihood of the proposition's truth, then updates his prior estimate in light of discoverable evidence bearing on that proposition, and arrives finally at a modified assessment of the proposition's likely truth in light of whatever evidence he has considered. When many items of evidence are involved, each has the effect of adjusting, in greater or lesser degree, the factfinder's evaluation of the probability that the proposition before him is true. If this ***1351** incremental process of cumulating evidence could be given quantitative expression, the factfinder might then be able to combine mathematical and non-mathematical evidence in a perfectly natural way, giving each neither more nor less weight than it logically deserves.

A quantitative description of the ordinary process of weighing evidence has long been available.[69] Before deciding whether that description can be put to the suggested use of enabling the factfinder to integrate mathematical and non-mathematical evidence, it will be necessary to develop the description briefly here.

Suppose *X* represents a disputed factual proposition; that is, the question for the trier is whether *X* is true or false. And suppose *E* represents some other proposition, the truth of which has just been established. Prior to learning *E,* the trier's subjective probability assessment of *X* was P(X). After learning *E,* the trier's assessment of *X* will typically change. That is, the trier's subjective probability for *X given the fact that E is true,* designated P(X|E),[70] will ordinarily differ from the trier's prior subjective probability for *X.*[71] The problem is to determine exactly how P(X|E), the probability of *X* given *E,* can be calculated in terms of P(X) and such other quantities as are available--to discover, that is, how the receipt of evidence *E* quantitatively transforms P(X) into P(X|E).

The solution to that problem, commonly known as Bayes' Theorem, will be summarized verbally after its mathematical formulation has been explained. The theorem follows directly from two elementary formulas of probability theory: if *A* and *B* are any two propositions, then:

(1) $P(A \& B) = P(A|B) \cdot P(B)$

(2) $P(A) = P(A \& B) + P(A \& \text{not-}B)$.[72]

***1352** These formulas can be shown to imply

(3) $P(X|E) = [P(E|X)/P(E)] \cdot P(X)$

and

(4) $P(E) = P(E|X) \cdot P(X) + P(E|\text{not-}X) \cdot P(\text{not-}X)$.[73]

And, using (4) to calculate P(E) in (3), we obtain

$$(5) \qquad P(X|E) \quad = \quad [P(E|X)/P(E|X) \bullet P(X) + P(E|\text{not-}X) \bullet P(\text{not-}X)] \bullet P(X).$$

Formula (5), known as Bayes' Theorem, determines P(X|E) in terms of P(X), P(E |X), and P(E|not-X).[74] In the abbreviated form of formula (3), Bayes' Theorem expresses the common sense notion that, to obtain P(X|E) from P(X), one multiplies the latter by a factor representing the probative force of *E*--that is, a factor equal to the ratio of P(E|X) (designating the probability of *E if X* is true) to P(E) (designating the probability of *E* whether or *not X* is true).[75]

***1353** Perhaps the easiest way to express Bayes' Theorem for the non-mathematician, though it is not the most convenient expression for actual use of the theorem, is to say that

$$(6)\ P(X|E) = P(E \,\&\, X)/P(E) = P(E|X) \bullet P(X)/P(E).$$

This simply asserts that the probability of *X* being true if *E* is known to be true, designated P(X|E), may be determined by measuring how often, out of all cases in which *E* is true, will *X also* be true--that is, by calculating the ratio of P(E & X) to P(E). That ratio, in turn, equals P(E|X) • P(X) divided by P(E), which completes the equation in Bayes' Theorem.

To give a concrete example, let *X* represent the proposition that the defendant in a particular murder case is guilty, and let *E* represent the evidentiary fact that the defendant left town on the first available plane after the murder was committed. Suppose that, prior to learning *E,* an individual juror would have guessed, on the basis of all the information then available to him, that the defendant was approximately twice as likely to be guilty as he was to be innocent, so that the juror's prior subjective probability for the defendant's guilt was P(X) = 2/3, and his prior subjective probability for the defendant's innocence was P(not-X) = 1/3. What effect should learning *E* have upon his probability assessment--*i.e.,* P(X|E)? The answer to that question will depend, of course, on how much more likely he thinks a guilty man would be to fly out of town immediately after the murder than an innocent man would be.

Suppose his best guess is that the probability of such flight if the defendant is guilty, designated P(E|X), is twenty percent, and that the probability of such flight if the defendant is innocent, designated P(E|not-X), is ten percent. Then P(E|X) = 1/5 and P(E|not-X) = 1/10. Recall formula (4):

$$P(E) = P(E|X) \bullet P(X) + P(E|\text{not-}X) \bullet P(\text{not-}X).$$

***1354** As applied to this case, we have

$$P(E) = (1/5)\,(2/3) + (1/10)\,(1/3) = 1/6.$$

In other words, given his prior assessment of P(X) = 2/3, the juror's best estimate of the probability of the defendant's flight, P(E), would have been 1/6. But if he knew that the defendant were in fact guilty, his best estimate of the probability that the defendant would flee, P(E|X), would be 1/5. Learning that he actually did flee should thus multiply the juror's prior assessment by the ratio

$$P(E|X)/P(E) = 1/5/1/6 = 6/5.$$

Applying formula (3),

$$P(X|E) = [P(E|X)/P(E)] \cdot P(X),$$

or

$$P(X|E) = [6/5] \cdot 2/3 = 4/5.$$

Therefore, his subsequent probability assessment of the defendant's guilt, after learning of his flight, should be 4/5. The evidence of flight should thus increase the juror's subjective probability estimate of the defendant's guilt from 2/3 to 4/5--assuming that he thinks there would be a 1/5 probability of flight if the defendant were guilty and a 1/10 probability of flight if he were not.

Given this precise a tool for cumulating evidence in a quantitative way, there might seem to be no obstacle to assimilating mathematical evidence into the trial process. Indeed, two commentators--one a lawyer, the other a statistician--have proposed doing exactly that. In an article in this *Review,* [76] Michael Finkelstein and William Fairley recently suggested that mathematically expert witnesses might be employed to explain to jurors the precise probative force of mathematical evidence in terms of its quantitative impact on the jurors' prior probability assessments. [77] As the next section of this article tries to show, although their analysis is both intriguing and illuminating, neither the technique proposed by Finkelstein and Fairley, nor any other ***1355** like it, can serve the intended purpose at an acceptable cost. It will be necessary first, however, to review the method they have suggested. To that end, it is useful to begin with the hypothetical case Finkelstein and Fairley posit.

A woman's body is found in a ditch in an urban area. There is evidence that the deceased quarreled violently with her boyfriend the night before and that he struck her on other occasions. A palm print similar to the defendant's is found on the knife that was used to kill the woman. Because the information in the print is limited, an expert can say only that such prints appear in no more than one case in a thousand. The question Finkelstein and Fairley ask themselves is how the jury might best be informed of the precise incriminating significance of that finding.

By itself, of course, the “one-in-a-thousand” statistic is not a very meaningful one. It does not, as the California Supreme Court in *Collins* showed, [78] measure the probability of the defendant's innocence-- although many jurors would be hard-pressed to understand why not. As Finkelstein and Fairley recognize, [79] even if there were as few as one hundred thousand potential suspects, one would expect approximately one hundred persons to have such prints; if there were a million potential suspects, one would expect to find a thousand or so similar prints. Thus the palm print would hardly pinpoint the defendant in any unique way.

To be sure, the finding of so relatively rare a print which matches the defendant's is an event of significant probative value, an event of which the jury should almost certainly be informed. [80] Yet the *numerical index* of the print's rarity, as measured by the frequency of its random occurrence, may be more misleading than enlightening, and the jury should be informed of that frequency--if at all--only if it is also given a careful explanation that there might well be many other individuals with similar prints. The jury should thus be made to understand that the frequency figure does not in any sense measure the probability of the defendant's innocence. [81]

***1356** Finkelstein and Fairley are distressed that this might leave the jury with too little information about the print's full probative value. The solution they propose to meet this difficulty is the use at trial of Bayes' Theorem; it is this solution to which I particularly object. [82] Let *X* represent the proposition that the defendant used the knife to kill his girlfriend, and let *E* represent the proposition that a palm print resembling the defendant's was found on the knife that killed her. P(E|X) is the probability of finding a palm print resembling the defendant's on the murder weapon if he was in fact the one who used the knife to kill, and P(E|not-X) is the probability of finding a palm print resembling the defendant's on the murder weapon if he was *not* the knife-user. P(X) represents the trier's probability assessment of the truth of *X* *before* learning *E,* and P(X|E) represents the trier's probability assessment of the truth of *X* *after* learning *E.* Finally, P(not-X) represents the trier's probability assessment of the falsity of *X* before learning *E,* so that P(not-X) = 1-P(X). Recall now Bayes' Theorem:

$$P(X|E) \quad = \quad [P(E|X)/P(E|X) \bullet P(X) + P(E|\text{not-}X) \bullet P(\text{not-}X)] \bullet P(X).$$

In applying this formula, Finkelstein and Fairley "assume for simplicity that defendant would inevitably leave such a print," [83] so that P(E|X) = 1. They also state that the probability P(E|not-X) equals "the frequency of the print in the suspect population." [84] In other words, they assume that the probability of finding a print like the defendant's on the knife, if the defendant did not in fact use the knife to kill his girlfriend, is equal to the probability that a randomly chosen person would have a print like the defendant's. In a later section of this article, [85] I will try to show that both of those assumptions are entirely unrealistic, that this error substantially distorts the results derived by Finkelstein and Fairley, and-- most importantly--that the error reflects not so much carelessness in their application of mathematical ***1357** methods as an inherent bias generated by the use of the methods themselves in the trial process. [86]

For now, however, my only purpose is to see where the method, as Finkelstein and Fairley apply it, leads us. They undertake, using Bayes' Theorem, to construct a table showing the resulting value of P(X|E) for a range of prior probabilities P(X) varying from .01 to .75, and for a range of different values for the frequency of the print in the suspect population varying from .001 to .50. Nine typical values for P(X|E) taken from the table Finkelstein and Fairley obtain (using the two simplifying assumptions noted above) are as follows: [87]

TABLE

POSTERIOR PROBABILITY OF *X* GIVEN *E* AS A FUNCTION OF FREQUENCY OF PRINT AND PRIOR PROBABILITY

	PRIOR PROBABILITY, P(X)		
FREQUENCY OF PRINT	**.01**	**.25**	**.75**
.50	.019	.400	.857
.10	.091	.769	.967
.001	.909	.997	.9996

The table shows, for example, that if a print like defendant's occurs with a frequency of one in a thousand, and if the trier's prior assessment of the probability that defendant used the knife to kill his girlfriend is one in four before he learns of the palm-print evidence *E,* then the palm-print evidence should increase to .997 the trier's posterior assessment of the probability that defendant used the knife to kill: P(X|E) = .997.

Finkelstein and Fairley would first have each juror listen to the evidence and arrive at a value for P(X) based upon his own view of the non-mathematical evidence (in this case, the prior quarrels and violent incidents). Then an expert witness would in effect show the jury the appropriate row from a table like the above, choosing the row to correspond with the testimony as to the print's frequency, so that each juror could locate the appropriate ***1358** value of P(X|E) as his final estimate of the probability that the defendant was in fact the knife-user. If the print's frequency were established to be .001, for example, the jurors need only be shown the last row; if it were .10, the jurors need only be shown the second row.[88] In this way, Finkelstein and Fairley argue, the frequency statistic would be translated for the jury into a probability statement which accurately describes its probative force. And, the authors add,[89] most of the respondents in an informal survey conducted by them would have derived higher final probabilities by this method than they did without the assistance of Bayes' Theorem. "Probably the greatest danger to a defendant from Bayesian methods," the authors conclude, "is that jurors may be surprised at the strength of the inference of guilt flowing from the combination of their prior suspicions and the statistical evidence. But this, if the suspicions are correctly estimated, is no more than the evidence deserves."[90]

Is it? We will be in a better position to answer that question at the end of the next section, which examines the costs we must be prepared to incur if we would follow the path Finkelstein and Fairley propose. What will presently be identified as certain costs of quantified methods of proof might conceivably be worth incurring if the benefit in increased trial accuracy were great enough. It turns out, however, that mathematical proof, far from providing any clear benefit, may in fact decrease the likelihood of accurate outcomes. It is the accuracy issue that I will consider first.

E. The Costs of Precision

1. The Distortion of Outcomes.--a. The Elusive Starting Point.--It is of course necessary, if the trier is to make any use at all of techniques like that proposed by Finkelstein and Fairley, for him first to settle on a numerical value for P(X), his assessment of the probability of *X* prior to the mathematical bombshell typically represented by the evidence *E.* But the lay trier will surely find it difficult at best, and sometimes impossible, to attach to P(X) a number that correctly represents his real prior assessment. Few laymen have had experience with the assignment of probabilities, and it might end up being a matter of pure chance whether a particular juror converts his mental state of partial certainty to a figure like .33, .43, or somewhere in between. An estimate of .5 might signify for one juror a guess ***1359** in the absence of any information and, for another, the conclusion of a search that has narrowed the inquiry to two equally probable suspects. And a juror's statement that he is "four-fifths sure," to revert to an earlier example,[91] is likely, in all but the simplest cases, to be spuriously exact.

Because the Finkelstein-Fairley technique thus compels the jury to begin with a number of the most dubious value, the use of that technique at trial would be very likely to yield wholly inaccurate, and misleadingly precise, conclusions. Even setting this threshold problem aside, major difficulties remain.

b. The Case of the Mathematical Prior.--Finkelstein and Fairley consider the application of their technique primarily to cases in which the prior probability assessment of a disputed proposition is based on non-mathematical evidence[92] and is then modified by the application of Bayes' Theorem to some further item of evidence that links the defendant to the case in a quantifiable way[93] or sheds some quantifiable light upon his conduct.[94] When statistical evidence is so used to modify a prior probability assessment, it is true, as the authors claim, that

> Bayesian analysis would demonstrate that the evidentiary weight of an impressive figure like one in a thousand--which might otherwise exercise an undue influence--would depend on the other evidence in the case, and might well be relatively insignificant if the prior suspicion were sufficiently weak.[95]

What they ignore, however, is that in most cases, whether civil or criminal, it will be none other than this "impressive figure like one in a thousand" that their general approach to proof would highlight. For, in most cases, the mathematical evidence will not ***1360** be such as to modify a prior probability assessment by furnishing added data about the specific case at hand. Instead, the mathematical evidence will typically bear only upon the broad category of cases of which the one being litigated will be merely an instance. Recall, for example, the situation in which sixty percent of all barrel-falling incidents were negligently caused, [96] the situations in which ninety-eight percent of all heroin was illegally imported, [97] the case in which four out of five blue buses belonged to the defendant, [98] or the prosecution in which the mistress was the murderess in ninety-five percent of all known similar instances. [99] In all of these cases, the mathematical evidence *E* simply describes the *class* of cases of which the litigated case is one. In such cases, *E* can *only* shed light on what initial value to assign to P(X). [100] Thus, the statistical information in these cases will, if given to the jury, create a high probability assessment of civil or criminal liability--and there is no assurance that the jury, either with or without the aid of Bayes' Theorem, will be able to make all of the adjustments in that high prior assessment that will be called for by the other evidence (or lack of it) that the rest of the trial reveals. The problem of the overpowering number, that one hard piece of information, is that it may dwarf all efforts to put it into perspective with more impressionistic sorts of evidence. This problem of acceptably combining the mathematical with the non-mathematical evidence is not touched in these cases by the Bayesian approach.

In situations of the sort being examined here, however, when the thrust of the mathematical evidence is to shed light on the probability assessment with which the trier ought rationally to begin, there is at least one way to take the evidence into account at trial without incurring the risk that the jury will give it too much weight when undertaking to combine the mathematical ***1361** datum with fuzzier information. Let the judge rather than the jury weigh the probabilistic proof in order to determine whether it might not be both equitable and conducive to an accurate outcome to shift to the other side the burden of producing some believable evidence to take the case outside the general rule seemingly established by the probabilities. [101] If one is to avoid a distortion in results, however, any such proposal must be qualified, at least when the question is one of the defendant's identity, by the principle that a party is not entitled to a jury verdict on statistical evidence alone absent some plausible explanation for his failure to adduce proof of a more individualized character. [102]

But the difficulty that calls forth this solution is not limited to cases in which the mathematics points to a readily quantifiable prior assessment. The problem--that of the overbearing impressiveness of numbers--pervades all cases in which the trial use of mathematics is proposed. And, whenever such use is in fact accomplished by methods resembling those of Finkelstein and Fairley, the problem becomes acute.

c. The Dwarfing of Soft Variables.--The syndrome is a familiar one: If you can't count it, it doesn't exist. Equipped with a mathematically powerful intellectual machine, even the ***1362** most sophisticated user is subject to an overwhelming temptation to feed his pet the food it can most comfortably digest. Readily quantifiable factors are easier to process--and hence more likely to be recognized and then reflected in the outcome--than are factors that resist ready quantification. The result, despite what turns out to be a spurious appearance of accuracy and completeness, is likely to be significantly warped and hence highly suspect.

The best illustration is none other than the computations performed by Finkelstein and Fairley themselves in their palm-print hypothetical. To begin with, they assume that, if the defendant had in fact used the knife to kill his girlfriend, then a palm print resembling his would certainly have been found on it, *i.e.,* P(E|X) = 1. [103] Had they not been moved by the greater ease of applying Bayes' Theorem under that assumption, the authors would surely have noted that a man about to commit murder with a knife might well choose to wear gloves, or that one who has committed murder might wipe off such prints as he happened to leave behind. Thus P(E|X) equals not 1, but 1-g, where *g* represents the combined probability of these contingencies and of the further contingency, later considered by Finkelstein and Fairley, [104] that such factors as variations within the suspected source [105] might prevent a print left by the defendant from seeming to match his palm. [106]

Far more significantly, Finkelstein and Fairley equate the frequency of the palm print in the suspect population with P(E|not-X), the probability of finding a print like the defendant's on the knife if he did not use it to kill. [107] That equation, however, strangely assumes that finding an innocent man's palm print on the murder weapon must represent a simple coincidence. If that were so, then the likelihood of such a coincidence, as measured by the print's frequency in the population, would of course yield the probability that a print like the defendant's would appear on the knife despite his innocence. [108] But this ignores the obvious fact that the print might have belonged to the defendant after all--without his having used the knife to kill the girl. He could simply have been framed, the real murderer having worn gloves when planting the defendant's knife at the ***1363** scene of the crime. The California Supreme Court recognized that sort of possibility in *Collins,* [109] noting that a jury traditionally weighs such risks in assessing the probative value of trial testimony. Finkelstein and Fairley, however, overlook the risk of frame-up altogether [110] --despite the nasty fact that the most inculpatory item of evidence may be the item most likely to be used to frame an innocent man. [111]

One can only surmise that it was the awkwardness of fitting the frame-up possibility into their formula that blinded even these sophisticated authors, one a legal scholar and the other a teacher of statistical theory, to a risk which they could not otherwise have failed to perceive. And if *they* were seduced by the mathematical machinery, one is entitled to doubt the efficacy of even the adversary process as a corrective to the jury's natural tendency to be similarly distracted. [112]

As it turns out, the frame-up risk would have been awkward indeed to work into the calculation, increasing P(E|not-X) from a value equal to the frequency of the palm print, hereinafter designated *f,* to a value equal to f + F, where *F* represents the ***1364** probability of frame-up. [113] Bayes' Theorem would then have assumed the messy form

$$P(X|E) = [(1-g)/(1-g)\ P(X) + (f + F)\ P(\text{not-}X)] \cdot P(X).$$ [114]

This formula may be easy enough to use when one assumes, with Finkelstein and Fairley, that $g = 0$ and $F = 0$, but is rather more troublesome to handle when one has no real idea *how* large those two probabilities are.

Moreover, it makes quite a difference to the outcome just how large the probabilities, *g* and *f,* turn out to be. Consider again the case in which the print is assumed to occur in one case in a thousand, so that *f* = .001, and in which the prior assessment is P(X) = .25. On those facts, Finkelstein and Fairley conclude-- by treating *g* and *F* as though they were both zero--that P(X |E) = .997, an overwhelming probability that the defendant did the killing. If, however, we assume that g = .1 and F = .1, then the same initial values f = .001 and P(X) = .25 yield the strikingly different conclusion P(X|E) = .75, a very much lower probability than Finkelstein and Fairley calculated. [115]

What, then, is one to tell the jurors? That each of them should arrive somehow at his own values for *g* and *F,* as to neither of which any evidence will be available, and should then wait while a mathematician explains what P(X|E) turns out to be given those values? Surely we have by now strained the system beyond its breaking point. [116] If the jurors are to work with mathematical proof in any capacity beyond that of passive observers, ***1365** they will in the end have to be given the numbers computed by Finkelstein and Fairley and asked to draw their own conclusions, keeping in mind--unless the judge who instructs them is as blinded by the formulas as the authors were-- that possibilities such as frame-up distort the figures in the table so that they overstate the truth by some indeterminate amount.

But then we have come full circle. At the outset some way of integrating the mathematical evidence with the non-mathematical was sought, so that the jury would not be confronted with an impressive number that it could not intelligently combine with the rest of the evidence, and to which it would therefore be tempted to assign disproportionate weight. At first glance, the use Finkelstein and Fairley made of Bayes' Theorem appeared to provide the needed amalgam. Yet, on closer inspection, their method too left a number--the exaggerated and much *more* impressive P(X|E) = .997-- which the jury must again be asked to balance against such fuzzy imponderables as the risk of frameup or of misobservation, if indeed it is not induced to ignore those imponderables altogether.

What is least clear in all of this is whether the proponents of mathematical proof have made any headway at all. Even assuming with Finkelstein and Fairley that the accuracy of trial outcomes could be somewhat enhanced if *all* crucial variables could be quantified precisely and analyzed with the aid of Bayes' Theorem, it simply does not follow that trial accuracy will be enhanced if *some* of the important variables are quantified and subjected to Bayesian analysis, leaving the softer ones--those to which meaningful numbers are hardest to attach--in an impressionistic limbo. On the contrary, the excessive weight that will thereby be given to those factors that can most easily be treated mathematically indicates that, on balance, more mistakes may well be made with partial quantification than with no quantification at all. [117]

d. Asking the Wrong Questions.--Throughout the preceding discussion, I have referred to P(X) and to P(X|E), deliberately eschewing the terminology employed by Finkelstein and Fairley. [118] Instead of *X,* they write *G*--impling that P(X|E) ***1366** represents the probability that the defendant is *guilty of murder* if a palm-print matching his is found on the murder weapon. But of course it represents no such thing, for murder means much more than causing death. To say that P(X|E) = .997 is to say that, given the palm-print evidence, there is a probability of .997 that the defendant used the knife to kill the deceased. It is to say nothing at all about his state of mind at the time, nothing about whether he intended to cause death, nothing about whether the act was premeditated. [119] To be sure, these elements can be called to each juror's attention, but his eyes are likely to return quickly to that imposing number on the board.

It is no accident that such matters as identity--matters that are objectively verifiable in the world outside the courtroom--lend themselves more readily to mathematical treatment than do such issues as intent--issues that correspond to no verifiable “fact” outside the verdict of the jury. It is not surprising that in none of the cases earlier enumerated under the heading of “intention” [120] was the mathematical evidence linked to any fact specifically about the defendant himself or about his own conduct or state of mind--for it is difficult even to imagine cases in which such a link could be found.

One consequence of mathematical proof, then, may be to shift the focus away from such elements as volition, knowledge, and intent, and toward such elements as identity and occurrence--for the same reason that the hard variables tend to swamp the soft. It is by no means clear that such marginal gains, if any, as we may make by finding somewhat more precise answers would not be offset by a tendency to emphasize the wrong questions. [121]

e. The Problem of Interdependence.--Essential to the application of Bayes' Theorem to derive P(X|E) from P(X) is that the trier be able somehow to make a prior estimate of P(X), the probability of the disputed proposition *X.* If that estimate is arrived at after knowing or even suspecting *E,* then to use the information provided by *E* to refine the estimate through Bayes' Theorem to obtain P(X|E) would obviously involve counting the same thing twice. Particularly when the proposition *X* goes to the identity of the person responsible for an alleged wrong, and when the tendency of *E* is to pinpoint the defendant as the ***1367** responsible person, the knowledge or suspicion of *E* is likely to have entered into the plaintiff's choice of this particular defendant. Having learned of the features of the interracial couple from the witnesses in *Collins,* [122] for example, the prosecution was hardly likely to charge someone *not* sharing those features; having found a latent palm print on the murder weapon, the State was less likely to file an indictment against a person whose palm print *failed* to match. And the trier would be hard put to disregard those obvious realities in attempting to derive a value for P(X). The accurate application of Bayes' Theorem along the lines proposed by Finkelstein and Fairley necessarily assumes that the evidence *E* of quantifiable probative value can be made independent of the prior suspicion, [123] but in most trials the two will be hopelessly enmeshed.

Indeed, even if P(X) is arrived at without any reliance whatever upon *E,* the straightforward application of Bayes' Theorem will still entail a distorted outcome if some or all of the evidence that *did* underlie P(X) was related to *E,* in the sense that knowing something about the truth of *X* and of that underlying evidence would yield information one way or the other about the likely truth of *E.* To take a simple example, suppose that an armed robbery taking fifteen minutes to complete was committed between 3.00 a.m. and 3:30 a.m. The trier first learns E_1: that the accused was seen in a car a half-mile from the scene of the crime at 3:10 a.m. Based on this information, the trier will assess a subjective probability P(X) of the accused's likely involvement in the robbery. Then the trier learns E_2: that the accused was also seen in a car a half-mile from the scene of the crime at 3:20 a.m. By itself, E_2 appears to make *X,* the proposition that the accused was involved, more likely than it would have seemed without any evidence as to the accused's whereabouts, so that $P(X|E_2)$ will exceed P(X) if computed by applying Bayes' Theorem directly, *i.e.,* by multiplying P(X) by the ratio $P(E_2|X)/P(E_2)$. Yet this is surely wrong, for if E_1 and E_2 are *both* true, then *X* must be false, and $P(X|E_2)$ should equal zero. [124] One ***1368** can with some effort make the appropriate adjustment--by using Bayes' Theorem to compute [125]

$$P(X|E_1 \& E_2) = [P(E_2|X \& E_1)/P(E_2|E_1)] \bullet P(X|E_1).$$

This means, however, that the theorem cannot be applied sequentially, with one simple multiplication by P(E|X) / P(E) as each new item of evidence, *E,* comes in, [126] but must instead be applied in the terribly cumbrous form shown above, [127] unless one knows somehow that the item of evidence to which the theorem is being applied at any given point is not linked conditionally [128] to any evidence already reflected in one's estimate of P(X). Finkelstein and Fairley ignore that requirement; taking it into account in order to avoid grossly inaccurate outcomes would make the machinery they propose so complex and so unwieldy that its operation, already hard enough for the juror to comprehend, would become completely opaque to all but the trained mathematician.

2. The End of Innocence: A Presumption of Guilt?--At least in criminal cases, and perhaps also in civil cases resting on allegations of moral fault, further difficulties lurk in the very fact that the trier is forced by the Finkelstein-Fairley technique to arrive at an explicit quantitative estimate of the likely truth at or near the trial's start, or at least before some of the most significant evidence has been put before him. [129]

To return for a moment to the palm-print case posited by Finkelstein and Fairley, a juror compelled to derive a quantitative measure P(X) of the defendant's likely guilt after having heard no evidence at all, or at most only the evidence about the ***1369** defendant's altercations with the victim, cannot escape the task of deciding just how much weight to give the undeniable fact that the defendant is, after all, not a person chosen at random but one strongly suspected and hence accused by officials of the state after extended investigation. If the juror undertakes to assess the probative value of that fact as realistically as he can, he will have to give weight to whatever information he supposes was known to the grand jury or to the prosecutor who filed the accusation. To the extent that this supposed information contains facts that will be duplicated by later evidence, such evidence will end up being counted twice, and will thus be given more weight than it deserves. [130] And, to the extent that the information attributed to the prosecutor or the grand jury is *not* duplicative in this sense, it will include some facts that will not be, and some facts that cannot properly be, introduced at trial as evidence against the accused. [131] Inviting jurors to take account of such facts is at war with the fundamental notion that the jury should make an independent judgment based only on the evidence properly before it, and would undercut the many weighty policies that render some categories of evidence legally inadmissible. [132]

***1370** Moreover, even if no such problem were present, directing the jury to focus on the probative weight of an indictment or other charge, or even directing it simply to assess the probability of the accused's guilt at some point before he has presented his case, would entail a significant cost. It may be supposed that no juror would be permitted to announce publicly in mid-trial that the defendant was already burdened with, say, a sixty percent probability of guilt--but even without such a public statement it would be exceedingly difficult for the accused, for the prosecution, and ultimately for the community, to avoid the explicit recognition that, having been forced to focus on the question, the rational juror could hardly avoid reaching some such answer. And, once that recognition had become a general one, our society's traditional affirmation of the “presumption of innocence” could lose much of its value.

That presumption, as I have suggested elsewhere, “represents far more than a rule of evidence. It represents a commitment to the proposition that a man who stands accused of crime is no less entitled than his accuser to freedom and respect as an innocent member of the community.” [133] In terms of tangible consequences for the accused, this commitment is significant because it can protect him from a variety of onerous restraints not needed to effectuate the interest in completing his trial; [134] because the suspension of adverse judgment that it mandates can encourage the trier to make an independent and more accurate assessment of his guilt; and because it may help to preserve an atmosphere in which his acquittal, should that be the outcome, will be taken seriously by the community. But no less important are what seem to me the intangible aspects of that commitment: its expressive and educative nature as a refusal to acknowledge prosecutorial omniscience in the face of the defendant's protest of innocence, and as an affirmation of respect for the accused--a respect expressed by the trier's willingness to listen to all the accused has to say before reaching any judgment, even a tentative one, as to his probable guilt.

It may be that most jurors would suspect, if forced to think about it, that a substantial percentage of those tried for crime ***1371** are guilty as charged. And that suspicion might find its way unconsciously into the behavior of at least some jurors sitting in judgment in criminal cases. But I very much doubt that this fact alone reduces the “presumption of innocence” to a useless fiction. The presumption retains force not as a *factual* judgment, but as a *normative* one--as a judgment that society *ought* to *speak* of accused men as innocent, and *treat* them as innocent, until they have been properly convicted after all they have to offer in their defense has been carefully weighed. The suspicion that many are *in fact* guilty need not undermine either this normative conclusion or its symbolic expression through trial procedure, so long as jurors are not compelled to articulate their prior suspicions of guilt in an explicit and precise way.

But if they *are* compelled to measure and acknowledge a factual presumption of guilt at or near each trial's start, then their underlying suspicion that such a presumption would often accord with reality may indeed frustrate the expressive and instructional values of affirming in the criminal process a normative presumption of innocence. Jurors cannot at the same time estimate probable guilt and suspend judgment until they have heard all the defendant has to say. It is here that the great virtue of mathematical rigor--its demand for precision, completeness, and candor--may become its greatest vice, for it may force jurors to articulate propositions whose truth virtually all might already suspect, but whose explicit and repeated expression may interfere with what seem to me the complex symbolic functions of trial procedure and its associated rhetoric.

To the extent that this argument and the one that immediately follows it [135] appear to run counter to rarely questioned assumptions about the transcending values of full candor and complete clarity, I should stress that the arguments in question are entirely independent of my other criticisms of mathematical methods in the trial process. Nor do I mean by advancing these arguments to suggest that departures from candor are lightly to be countenanced. Indeed, I have not proposed that anyone deceive either himself or another about the factual underpinnings of the presumption of innocence, but only that worthwhile values served by that presumption as a normative standard might be harder to secure if the probability of guilt became a matter for precise and explicit assessment and articulation early in the typical criminal trial. The point, then, is not that any factual truth should be concealed or even obscured, but only that one need not say everything all at once in order to be truthful, and that saying some things in certain ways and at certain times in the trial process ***1372** may interfere with other more important messages that the process should seek to convey and with attitudes that it should seek to preserve.

3. The Quantification of Sacrifice.--This concern for the expressive role of trial procedure is no less relevant to the trial's end than to its start. Limiting myself here to the ordinary criminal proceeding, [136] I suggest that the acceptance of anything like the method Finkelstein and Fairley propose, given the precision and explicitness its use demands, could dangerously undermine yet another complex of values--the values surrounding the notion that juries should convict only when guilt is beyond real doubt. [137]

An inescapable corollary of the proposed method, and indeed of any method that aims to assimilate mathematical proof by quantifying the probative force of evidence generally, is that it leaves the trier of fact, when all is said and done, with a number that purports to represent his assessment of the probability that the defendant is guilty as charged. [138] Needless to say, that number will never quite equal 1.0, so the result will be to produce a quantity--take .95 for the sake of illustration--which openly signifies a measurable (and potentially reducible) margin of doubt, here a margin of .05, or 1/20.

Now it may well be, as I have argued elsewhere, [139] that there is something intrinsically immoral about condemning a man as a criminal while telling oneself, “I believe that there is a chance of one in twenty that this defendant is innocent, but a 1/20 risk of sacrificing him erroneously is one I am willing to run in the interest of the public's--and my own--safety.” It may be that ***1373** --quite apart from the particular number--there is something basically immoral in this posture, but I do not insist here on that position. All I suggest is that a useful purpose may be served by structuring a system of criminal justice so that it can avoid having to proclaim, as the Finkelstein-Fairley procedure would force us to proclaim, that it will impose its sanctions in the face of a recognized and quantitatively measured doubt in the particular case.

If the system in fact did exactly that, such compelled candor about its operation might have great value. It could generate pressure for useful procedural reform, and it should probably be considered worthwhile in itself. [140] But to let the matter rest ***1374** there would be wrong, for the system does *not* in fact authorize the imposition of criminal punishment when the trier recognizes a quantifiable doubt as to the defendant's guilt. Instead, the system dramatically--if imprecisely--insists upon as close an approximation to certainty as seems humanly attainable in the circumstances. [141] The jury is charged that any “reasonable doubt,” of whatever magnitude, must be resolved in favor of the accused. Such insistence on the greatest certainty that seems reasonably attainable can serve at the trial's end, like the presumption of innocence at the trial's start, [142] to affirm the dignity of the accused and to display respect for his rights as a person--in this instance, by declining to put those rights in deliberate jeopardy and by refusing to sacrifice him to the interests of others.

In contrast, for the jury to announce that it is prepared to convict the defendant in the face of an acknowledged and numerically measurable doubt as to his guilt is to tell the accused that those who judge him find it preferable to accept the resulting risk of his unjust conviction than to reduce that risk by demanding any further or more convincing proof of his guilt. I am far from persuaded that this represents the sort of thought process through which jurors do, or should, arrive at verdicts of guilt. Many jurors would no doubt describe themselves as being "completely sure," or at least as being "as sure as possible," before they vote to convict. That some mistaken verdicts are inevitably returned even by jurors who regard themselves as "certain" is of course true but is irrelevant; such unavoidable errors are in no sense *intended,* [143] and the fact that they must occur if trials are to be conducted at all need not undermine the effort, through the symbols of trial procedure, to express society's fundamental commitment to the protection of the defendant's rights as a person, as an end in himself. On the other hand, formulating an "acceptable" risk of error to which the trier is willing deliberately to subject the defendant would interfere seriously with this expressive role of the demand for certitude--however unattainable real certitude may be, and however clearly all may ultimately recognize its unattainability.

***1375** In short, to say that society recognizes the necessity of tolerating the erroneous "conviction of some innocent suspects in order to assure the confinement of a vastly larger number of guilty criminals" [144] is not at all to say that society does, or should, embrace a policy that juries, *conscious of the magnitude of their doubts in a particular case,* ought to convict in the face of this acknowledged and quantified uncertainty. It is to the complex difference between these two propositions that the concept of "guilt beyond a reasonable doubt" inevitably speaks. The concept signifies not any mathematical measure of the precise degree of certitude we require of juries in criminal cases, [145] but a subtle compromise between the knowledge, on the one hand, that we cannot realistically insist on acquittal whenever guilt is less than absolutely certain, and the realization, on the other hand, that the cost of spelling that out explicitly and with calculated precision in the trial itself would be too high. [146]

4. The Dehumanization of Justice.--Finally, we have been told by Finkelstein and Fairley that jurors using their method may find themselves "surprised" at the strength of the inference of guilt flowing from the combination of mathematical and nonmathematical evidence. [147] Indeed they may, [148] and in a far deeper sense than with other equally obscure forms of expert testimony, for such testimony typically represents no more than an input into the trial process, whereas the proposed use of Bayesian methods changes the character of the trial process itself. When that change yields a "surprisingly" strong inference of guilt in a particular case, it is by no means clear that, so long as one keeps one's numbers straight, "this … is no more than the evidence deserves." [149] Methods of proof that impose moral ***1376** blame or authorize official sanctions [150] on the basis of evidence that fails to penetrate or convince the untutored contemporary intuition threaten to make the legal system seem even more alien and inhuman than it already does to distressingly many. There is at stake not only the further weakening of the confidence of the parties and of their willingness to abide by the result, but also the further erosion of the public's sense that the law's fact-finding apparatus is functioning in a somewhat comprehensible way, on the basis of evidence that speaks, at least in general terms, to the larger community that the processes of adjudication must ultimately serve. The need now is to enhance community comprehension of the trial process, not to exacerbate an already serious problem by shrouding the process in mathematical obscurity.

It would be a terrible mistake to forget that a typical lawsuit, whether civil or criminal, is only in part an objective search for historical truth. It is also, and no less importantly, a ritual--a complex pattern of gestures comprising what Henry Hart and John McNaughton once called "society's last line of defense in the indispensable effort to secure the peaceful settlement of social conflicts." [151]

One element, at least, of that ritual of conflict-settlement is the presence and functioning of the jury--a cumbersome and imperfect institution, to be sure, but an institution well calculated, at least potentially, to mediate between "the law" in the abstract and the human needs of those affected by it. Guided and perhaps intimidated by the seeming inexorability of numbers, induced by the persuasive force of formulas and the precision of decimal points to perceive themselves as performing a largely mechanical and automatic role, few jurors--whether in criminal cases or in civil--could be relied upon to recall, let alone to perform, this humanizing function, to employ their intuition and their sense of community values to shape their ultimate conclusions. [152]

When one remembers these things, one must acknowledge that there was a wisdom of sorts even in trial by battle--for at least that mode of ascertaining truth and resolving conflict reflected ***1377** well the deeply-felt beliefs of the times and places in which it was practiced. [153] This is something that can hardly be said of trial by mathematics today.

F. Conclusions

I am not yet prepared to say that the costs of mathematical precision enumerated here are so great as to outweigh any possible gain that might be derived from the carefully limited use of probabilistic proof in special circumstances. I do think it clear, however, that those circumstances would have to be extraordinary indeed for the proponents of mathematical methods of proof to make even a plausible case.

With the possible exception of using statistical data to shift the burden of production, [154] and perhaps with the further exception of using evidence as to frequencies in order to negate a misleading impression of uniqueness that expert opinion might otherwise convey, [155] I think it fair to say that the costs of attempting to integrate mathematics into the factfinding process of a legal trial outweigh the benefits. In particular, the technique proposed by Finkelstein and Fairley is incapable of achieving ***1378** the objectives claimed for it, and possesses grave deficiencies that any other similarly conceived approach would be very likely to share.

It does not follow, however, that mathematical methods must play an equally limited role in the enterprise of designing trial procedures. What is true of mathematics as an aid to factfinding may be false of mathematics as an aid to rulemaking. In the pages that follow, I examine this separate issue and attempt to show that, although one can have somewhat more hope for mathematics in rulemaking, special problems of a quite serious character arise in that related context as well.

II. RULEMAKING WITH NUMBERS AND CURVES

A. One Simplified Model

Thus far, I have considered the role of mathematics in the process of proof, its potentialities and limitations in helping the trier of fact assess the probability of a disputed proposition. Once that probability has been assessed--by whatever means, mathematical or otherwise--there remains the problem of deciding what to do, what verdict to return. Can mathematical techniques be of assistance in formulating a rule of procedure--a standard of proof--that will solve that problem? More generally, what role can mathematics play in designing procedural rules for the trial process?

I want to consider first the narrower of those two questions, for it is the only one to which significant effort has thus far been directed. John Kaplan [156] and Alan Cullison [157] have both proposed a rather simple mathematical model of the trial process in order to determine the probability necessary to return a verdict. Although the model is as applicable to civil cases as to criminal, it is most readily understood in the setting of a criminal trial. They propose that a criminal trial be viewed as analogous to any other situation in which one must choose between two or more courses of action on the basis of a body of information which reduces, but does not wholly eliminate, the decisionmaker's uncertainty about the true state of the world and about the consequences in that world of any chosen strategy of conduct. [158]

In particular, the trier must choose between conviction (designated *C*) and acquittal (designated *A*) in the face of at least partial uncertainty as to whether the defendant is in fact guilty ***1379** (designated *G*) or innocent (designated *I*). The four possible outcomes of the trier's decision problem are:

Outcome

(1) Convicting a guilty man, designated C_G

(2) Convicting an innocent man, designated C_I

(3) Acquitting a guilty man, designated A_G

(4) Acquitting an innocent man, designated A_I

The model posits that the rational trier should [159] choose *C* rather than *A* whenever the "expected utility" to the trier of the former choice would exceed the "expected utility" of the latter, in light of such factors as the seriousness of the offense, the severity of the punishment, and so on--much as a rational gambler would select the bet that maximizes his expected gains, taking into account his present position, his needs, and his attitudes toward risk. [160]

In order to make the necessary choice, the trier must first decide how much he would like or dislike each of the four possible outcomes of the proceeding-- that is, he must decide what "utility" each has for him. [161] Suppose that the trier's order of preference, from the outcome he would like best to the one he would like least, is C_G, A_I, A_G, C_I. In order to assign quantitative utilities $U(C_G)$, $U(A_I)$, $U(A_G)$, and $U(C_I)$ to these outcomes, he begins by assigning a maximum utility of 1 to the outcome he likes most and a minimum utility of 0 to the outcome he likes least:

$\mathbf{U(C_G) = 1}$

$\mathbf{U(C_I) = 0}$

To decide what utility between 0 and 1 to assign A_I, the trier ***1380** asks himself such questions as the following: Would I rather get A_I for sure or get a 1/2 chance of the best outcome, C_G, and a 1/2 chance of the worst, C_I? If the answer is that he would rather get A_I, then $U(A_I)$ is said to exceed 1/2; if he would prefer the gamble, $U(A_I)$ is less than 1/2. If it turns out that $U(A_I)$ exceeds 1/2 by this test, then the trier asks himself whether he would rather get A_I for sure or get, say, a 3/4 chance of C_G and a 1/4 chance of C_I. If the answer this time is that he would rather take the chance, then $U(A_I)$ falls between 1/2 and 3/4. In this way, the trier "closes in" on $U(A_I)$ until he ultimately pinpoints its value. To say, for example, that $U(A_I) = 2/3$ is to say that the trier would be as satisfied getting A_I for sure as he would be getting a 2/3 chance of C_G and 1/3 chance of C_I. Suppose for the sake of illustration that, by this same process, the trier concludes that, for him, $U(A_G) = 1/2$.

Now the trier is in a position to decide how sure of the defendant's guilt he would have to be before preferring *C* to *A.* To that end, let *P* designate the trier's probability assessment of *G* in light of all the evidence in the case. Then, if the trier chooses *C,* there is a probability of *P* that he will get C_G and a probability of 1-P that he will get C_I. EU(C), the "expected utility" or "expected desirability" of this choice, is the sum of two products: (1) the probability of guilt, *P,* multiplied by the desirability of C_G; and (2) the probability of innocence, 1-P, multiplied by the desirability of C_I:

$$EU(C) = P \bullet U(C_G) + (1-P) \bullet U(C_I).$$

But this simply equals *P,* since $U(C_G) = 1$ and $U(C_I) = 0$. If the trier chooses *A,* there is a probability of *P* that he will get A_G and a probability of 1-P that he will get A_I. EU(A), the "expected utility" of *A,* is thus

$$EU(A) = P \bullet U(A_G) + (1-P) \bullet U(A_I),$$

which in our case equals $P \bullet 1/2 + (1-P) \bullet 2/3$, or 4-P/6. Thus the expected utility of choosing *C* exceeds that of choosing *A* whenever *P* exceeds 4-P/6, which occurs whenever *P* exceeds 4/7.

Hence, given the utilities the trier has assigned to the four possible outcomes, the model supplies him with a rule of procedure for this criminal case: "Consider the evidence and then vote to convict if and only if you think that the probability of the defendant's guilt exceeds 4/7." [162]

***1381** Now, one might have qualms about the resulting procedural rule-- because one regards a threshold probability of 4/7 as much too low, or because one objects in principle to the willing taking of *any* measurable risk of convicting an innocent man, [163] or because one regards as unacceptable the cost of openly *announcing* that willingness [164] --but my concern here is not so much with the result as with the method used to arrive at it. It is to a criticism of that method that the next section is addressed.

B. A Critique of the Model

1. *Misspecification of Consequences.*--The model described above assumes the existence of meaningful answers to such questions as: "How much would you regret the erroneous conviction of this defendant for armed robbery?" But the answer must surely be "It depends." It depends in part upon the *character* of the error itself; mistaken identity might be worse, for instance, than misjudged intention and worse still than a miscalculated statute of limitations. [165] And it depends even more significantly upon the *process* that led to the error; one cannot equate the lynching of an innocent man with his mistaken conviction after a fair trial. Indeed, it is at least arguable that there is nothing good or bad about *any* trial outcome as such; that the *process,* and not the *result in any particular case,* is all-important. To be sure, some concern for the mix of correct and erroneous outcomes operates as a constraint on what might otherwise ***1382** be

deemed acceptable trial procedures--but the acceptability of a process is not simply a function of the number of correct or erroneous convictions or acquittals it yields. At the very least it is clear that our preferences, and those of the trier, attach not to the bare consequences of correct or erroneous conviction or acquittal. They attach instead, and properly so, to the consequences--for a broad range of values and interests--of the defendant's correct or erroneous conviction or acquittal after a given sort of trial, operating with a particular set of rules and biases, and governed by a specific standard of proof. [166]

In particular, the trier might justly regard as worse the erroneous conviction of a man to whose guilt he had attached a probability of just over 4/7 than the erroneous conviction of one whose guilt had seemed to be virtually certain. Indeed, the trier would probably have to attach a different "utility" to the outcome of erroneously convicting a man on the basis of a standard that appeared to convey to the community at large a willingness to take calculated risks of such errors, than he would to the outcome of erroneously convicting a man on the basis of a standard that gave no such appearance. [167] At a minimum, therefore, because the utilities of the various consequences would themselves be functions of the apparent probability of the defendant's ***1383** guilt, any equation designed to compute the threshold probability above which conviction would be preferable to acquittal would have to be far more complex than Kaplan and Cullison have supposed. [168] And that, in turn, could preclude the existence of *any* single threshold and would in any event make the model, already too obscure for actual use by a trier of fact, more esoteric still.

2. Problems of Cognitive Dissonance.--The preceding discussion demonstrates that a legal trial differs from the usual sort of management problem to which utility theory has previously been applied [169] in at least one important respect: various features of the procedure followed to reach the decision, including the standard of proof applied, are themselves integral parts of the consequences to be optimized, a fact that greatly complicates the optimization process. The trial decision also differs from the classical management problem in another crucial respect: the decisionmaker will invariably have preferences not only with respect to the consequences of his choice but also with respect to the underlying facts themselves, facts over which he can exercise no control. Thus, for example, the trier's reluctance to see an innocent man put to the ordeal of trial and his wish to avoid discovering that the man who is in fact guilty remains at liberty may combine to reduce the desirability, for him, of the outcome previously designated A_I (acquittal of the defendant who is in fact innocent), as compared with the outcome previously designed C_G (conviction of the defendant who is in fact ***1384** guilty). [170] The risk is that the trier will not only allow his hope that the accused is in fact guilty to influence his perception of the evidence--a classic case of adjusting cognition to avoid psychological dissonance [171] --but will also allow that hope, through a distortion in the comparative magnitudes of U(A_I) and U(C_G), to influence his determination of what standard of proof to apply. The suggested method for arriving at that standard in no way guards against this danger.

3. Positing the Wrong Decisionmaker.--It may be that *no* method of arriving at the standard of proof can avoid the problem identified above unless it effectively separates the selection of that standard from the decision of the particular case. Indeed, a variety of other important values, including that of both real and apparent equality in the treatment of accused persons, all point in the same direction: the factfinder in a criminal trial should not be encouraged to do what the proposed model demands of him--namely, that he expressly assess the desirability or "utility" of the various consequences that might flow from correctly or erroneously convicting or acquitting the particular defendant then on trial. [172]

The fact that such matters as the defendant's reputation and the likely sentence, all of which, of course, bear directly on the utilities of the four possible outcomes, are nonetheless often kept from the jury "is hard to defend … on a decision-theoretic view" [173] only if one's decision theory neglects to ask about the institutional competencies of the several elements of the legal system. But when one gives due weight to the costs of combining in the trier the separate functions of deciding what happened in a particular case and evaluating the anticipated consequences of alternative verdicts, one will expect the lawmaker rather than the factfinder to use a model such as the one Kaplan and Cullison propose, and one will define the decision problem to be ***1385** solved not as the one-shot problem of fixing a standard of proof for a particular trial with four possible outcomes, but as the much larger problem of establishing such standards for the trial system as a whole.

4. Operating in a Factual Vacuum.--Having thus broadened the inquiry, one cannot avoid noticing that the model proposed by Kaplan and Cullison is, oddly enough, structured to make use of *none* of the crucial facts one would surely want to know when establishing a standard of proof. We are, after all, talking about some real things: crimes, their prevention, the incapacitation of their perpetrators, and the protection of innocent persons from being falsely convicted for them. If it be proposed that juries in a certain kind of case should convict whenever they think a defendant's probability of guilt exceeds 4/7, no one concerned with these real things could fail to ask questions such as these: How many guilty men are likely to be erroneously acquitted under that standard? How many innocent men are likely to be erroneously convicted? What will be the effect on the likely number of offenses? What will be the impact on the fear of false prosecution and unjust imprisonment? The answers to such questions, in turn, will depend on such other inquiries as these: How easy or difficult is it for the state to make the probability of an innocent man's guilt appear to exceed 4/7? How easy is it for a guilty defendant to make the probability of his innocence appear to exceed 3/7? How many innocent men are brought to trial? How might the number of innocents tried depend on the announced standard of proof? How do the number of offenses or the fear of erroneous conviction relate to the probability of conviction if guilty? To the probability of conviction if innocent? To the ratio of convictions to acquittals? To the absolute number of convictions? To the absolute number of acquittals?

The striking thing is that the answers to virtually none of these obviously relevant questions could ever find their way into the Kaplan-Cullison model, for the answers simply do not relate, in the main, to an assessment of the desirability or undesirability of one or another outcome of a paradigm trial; they relate instead to the characteristics of a much broader system.[174] The proposed model, then, is not really useful and does not provide ***1386** a fair test for the potentialities of mathematical methods in procedural design.

C. More Sophisticated Techniques

A somewhat fairer test might be provided by an approach employing what economists usually call “choice sets” and either “indifference curves” or “preference contours.”[175] For a particular crime, the rulemaker would first establish the “choice set” open to him by investigating the functional relationship to be expected between the percentage of guilty convicted and the percentage of innocent convicted, recognizing that--at any given level of resource investment--convicting more of the guilty may require relaxing various procedural standards (including, but not limited to, the probability of guilt required for conviction in any particular case) and thereby convicting more of the innocent as well. That functional relationship will reflect, among other things, the ratio of guilty to innocent defendants among those brought to trial for the crime in question, the sensitivity of trial outcomes to various procedural rules, and a number of other factors that would surely vary from one jurisdiction to another, and from one crime to another.

Having thus established in some empirical way the choices open to him with respect to this crime, the rulemaker would next think about his preferences, or those of his constituents. If he could convict, for example, 60% of the guilty with a 1% chance of convicting an innocent, how much would he let the latter percentage rise in order to convict an additional 5% of the guilty? To convict an additional 10% of the guilty? In thinking about the answers to such questions, the rulemaker would, of course, have to take account of whatever information he could develop on such topics as the relationship between the probability of conviction if guilty and the corresponding frequency of offenses,[176] and he would also have to take account of such factors as the relationship between the probability of conviction if innocent ***1387** and the corresponding level of fear and insecurity among his constituents.[177]

Starting with any arbitrarily chosen point--such as the one at which 60% of the guilty and 1% of the innocent are convicted--the rulemaker would consider such relationships in deriving his "preference contour" through the point (60, 1):

TABULAR OR GRAPHIC MATERIAL SET FORTH AT THIS POINT IS NOT DISPLAYABLE

The curve drawn indicates, among other things, that the rule-maker would be willing to let the figure of 1 climb to 1.5 (but no higher) in order to increase the 60 to 65, and to 2 (but no higher) in order to increase the 65 to 70. It also indicates that he would be willing to let the 60 drop to 56 (but no lower) in order to decrease the 1 to .5.

The rulemaker would then draw another preference contour starting at the point (60, 2); another one starting at (60, 3); ***1388** and so on, thereby building up a complete set of indifference curves or preference contours:[178]

TABULAR OR GRAPHIC MATERIAL SET FORTH AT THIS POINT IS NOT DISPLAYABLE

Finally, the rulemaker would superimpose on these preference contours the empirically established "choice set" *C*--the functional relationship telling him, at the assumed level of resource investment, how many innocents would be convicted for any given percentage of guilty:

TABULAR OR GRAPHIC MATERIAL SET FORTH AT THIS POINT IS NOT DISPLAYABLE

The optimum point on the choice set is *Q,* the point of tangency between that set and a preference contour.[179] By finding ***1389** that point of tangency, the rulemaker determines the percentage of guilty convictions he should aim for and knows the corresponding percentage of innocent convictions that will result. If, for example, *Q* is at the point (80, 1.2), the rulemaker knows that he should design the procedure for trials of the crime in question so as to convict some 80% of the guilty at a cost of convicting some 1.2% of the innocent. The problem of discovering just what combination of procedures (standards of proof, presumptions, rules of admissibility and exclusion, and the like) will have that approximate effect then becomes the next task-- obviously a difficult one--on the rulemaker's agenda for empirical research and mathematical analysis.

D. Some Tentative Reservations

1. On Precision and Quantification.--To whatever extent all of this represents, albeit in somewhat simplified and preliminary form, a typical instance of mathematical reasoning in the design of trial procedures, it is important to explore the costs that may be incurred by its use.

Those costs, in large measure, are the same costs of precision that I have examined in another context.[180] In particular, there is a significant risk that the greater ease with which the rule-maker will be able to quantify some variables (such as the incidence of crime) as compared to others (such as the insecurity flowing from fear of false conviction) will skew his decision in unfortunate directions, leaving serious doubt whether the exactitude of the numbers and curves will, in the end, lead to better rules.[181] In the preference contour exercise attempted above, for example, it seems clear that the shape of the contours ought at least to reflect the procedures and the philosophy that would be required to achieve any given mix of trial outcomes. One cannot really say, for example, whether one feels better or worse about convicting

***1390** 80% of the guilty and 1.2% of the innocent than one feels about convicting 90% of the guilty and 2.5% of the innocent, unless one knows how trial procedures might have to be altered [182] in order to go from the former point to the latter. But the costs of that procedural alteration, in terms of the many intangible consequences of such a change for a broad spectrum of values, will almost certainly prove harder to quantify than will the benefits of convicting another 10% of the guilty. [183] As a result, the preference contours may fail to reflect how the rule-maker really *does* feel about things--and the conclusion to which they point may be less acceptable than one more intuitively and impressionistically derived.

Moreover, once one is precise and calculating about rule-making, one can no longer so easily enjoy the benefits of those profoundly useful notions--like the "presumption of innocence" and "acquittal in all cases of doubt"--that we earlier saw threatened by mathematical proof. [184] After deciding in a deliberate and calculated way that it is willing to convict twelve innocent defendants out of 1000 in order to convict 800 who are guilty--because that is thought to be preferable to convicting just six who are innocent but only 500 who are guilty--a community would be hard pressed to insist in its culture and rhetoric that the rights of innocent persons must not be deliberately sacrificed for social gain. [185]

There are, finally, several problems of a different order-- problems that go to the wisdom of being somewhat fuzzy and open-ended in one's statement of at least some kinds of standards and procedures that are designed to guide others over time. I have in mind the great advantage in some areas of principles over rules, [186] of formulations that facilitate consensus on results ***1391** [187] and leave one free to move in many different directions as one's understanding grows and as one's needs evolve. [188] There is no necessary reason why mathematical analysis, operating with deliberately unspecified variables, cannot someday prove helpful in this subtle business--but I doubt that the day has come. At least in the rudimentary state of the art represented by the preceding two sections, the mission of mathematics *is* specification, and the almost inevitable corollary of its serious use in these circumstances is a move away from the open-ended to the rigorously defined.

2. On Utility and Ritual.--The appropriateness of applying mathematical methods to decisionmaking seems clearest when the alternative acts among which one is deciding are significant only as means to some external set of agreed-upon ends. For the decisionmaker can then approach his problem as the essentially mechanical one of choosing the act whose expected consequences will maximize a suitably weighted combination of those ends, subject to some appropriately defined set of constraints.

The great difficulty with thinking in this way about the choice of legal rules and the design of legal institutions is that such rules and institutions are often significant, not only as means of achieving various ends external to themselves, but also as ends in their own right, or at least as symbolic expressions of certain ends and values.

As much of the preceding analysis has indicated, [189] rules of trial procedure in particular have importance largely as expressive entities and only in part as means of influencing independently significant conduct and outcomes. [190] Some of those rules, to be sure, reflect only "an arid ritual of meaningless form," [191] but others express profoundly significant moral relationships and principles--principles too subtle to be translated into anything less complex than the intricate symbolism of the trial process. Far from being either barren or obsolete, much of what goes on in the trial of a lawsuit--particularly in a criminal case--is partly ceremonial or ritualistic in this deeply positive sense, and partly educational as well; procedure can serve a vital role as conventionalized communication among a trial's participants, and as something like a reminder to the community of the principles ***1392** it holds important. [192] The presumption of innocence, [193] the rights to counsel [194] and confrontation, [195] the privilege against self-incrimination, [196] and a variety of other trial rights, [197] matter not only as devices for achieving or avoiding certain kinds of trial outcomes, [198] but also as affirmations of respect for the accused as a human being--affirmations that remind him and the public about the sort of society we want to become and, indeed, about the sort of society we are. [199]

Perhaps these expressive roles of procedure can be formally assimilated into a utility-maximizing model by adding on appropriate values to the weighted combination of preferred ends. [200] But, however completely this amplification of the model mirrors all of one's values, there is little chance of capturing the fact that much of what matters about expressive rules, procedural or otherwise, is that they *embody* and do not merely *implement* the values of the community that follows them. To employ mathematical techniques to help choose that rule which will maximize an appropriately weighted mix of certain values or preferences is to take those values as given--as objects outside the rules among which one is choosing. In fact, however, the very *choice* of one rule rather than another--of a rule that the accused cannot be forced to testify against himself, for example--may itself evidence and indeed constitute a *change* in the mix of basic values of the society that has made the choice in question. [201] At this point, the decision problem--if it can ***1393** still be called that--is to "choose" what fundamental values one wants to have and not simply to find the best way of implementing a set of values accepted as given. [202] Numbers and curves can be of relatively little use at so ultimate a level.

E. Conclusions

Reluctant as I am to make confident pronouncements about the final limits of mathematics in the fact-finding process of a civil or criminal trial, [203] I am more reluctant still to attempt any definitive assessment of how far mathematical methods and models can acceptably be exploited in the rulemaking process that determines how trials are conducted.

I have examined in some detail one simple model proposed by Kaplan and Cullison to assist in the determination of standards of proof, and have concluded that their approach, like that of Finkelstein and Fairley in the context of mathematical evidence, is more misleading than helpful. I have analyzed less closely the outlines of a more complex methodology--one that would apply preference contours and choice sets to the derivation of rules of criminal procedure--and have found that methodology substantially more enlightening but still far from satisfactory. And I have attempted to show, finally, that there may be at least some inherent limitations in the linking of mathematics to procedural rulemaking--limitations arising in part from the tendency of more readily quantifiable variables to dwarf those that are harder to measure, in part from the uneasy partnership of mathematical precision and certain important values, in part from the possible incompatibility of mathematics with openended and deliberately ill-defined formulations, and in part from the intrinsic difficulty of applying techniques of maximization to the rich fabric of ritual and to the selection of ends as opposed to the specification of means.

In an era when the power but not the wisdom of science is increasingly taken for granted, there has been a rapidly growing interest in the conjunction of mathematics and the trial process. The literature of legal praise for the progeny of such a wedding has been little short of lyrical. Surely the time has come for someone to suggest that the union would be more dangerous than fruitful.

Footnotes

[a1] Assistant Professor of Law, Harvard University. A.B. Harvard, 1962; J.D. Harvard, 1966.

[1] *See* M. CAPPELLETTI & J. PERILLO, CIVIL PROCEDURE IN ITALY 35-36 (1965); A. ENGELMANN, A HISTORY OF CONTINENTAL CIVIL PROCEDURE 41-47 (1927); R. GINSBURG & A. BRUZELIUS, CIVIL PROCEDURE IN

SWEDENNN 33 & n.131, 295 & n.471 (1965); J. GLASER, LEHRE VOM BEWEIS IM STRAFPROZESS 132-35 (1883); Kunert, *Some Observations on the Origin and Structure of Evidence Rules Under the Common Law System and the Civil Law System of "Free Proof" in the German Code of Criminal Procedure,* 16 BUFF. L. REV. 122, 141-42 & nn.99-100, 144-45 (1966). *See also* A. ESMEIN, A HISTORY OF CONTINENTAL CRIMINAL PROCEDURE 264-71 (J. Simpson, transl. 1913); 1 F. HÉLIE, TRAITÉ DE L'INSTRUCTION CRIMINELLE 650-53, 656-57 (1845); F. VOLTAIRE, A COMMENTARY ON BECCARIA'S ESSAY ON CRIMES AND PUNISHMENTSSSS 227-28 (1872).

[2] I am, of course, aware that *all* factual evidence is ultimately "statistical," and all legal proof ultimately "probabilistic," in the epistemological sense that no conclusion can ever be drawn from empirical data without some step of inductive inference--even if only an inference that things are usually what they are perceived to be. *See, e.g.,* D. HUME, A TREATISE OF HUMAN NATURE, bk. I, pt. III, § 6, at 87 (L.A. Selby-Bigge ed. 1958). My concern, however, is only with types of evidence and modes of proof that bring this probabilistic element of inference to explicit attention in a quantified way. As I hope to show, much turns on whether such explicit quantification is attempted.

[3] By "mathematical methods," I mean the entire family of formal techniques of analysis that build on explicit axiomatic foundations, employ rigorous principles of deduction to construct chains of argument, and rely on symbolic modes of expression calculated to reduce ambiguity to a minimum.

[4] One senses that much of the contemporary opposition to the technological emphasis upon rationality and technique rests on some such premise.

[5] *See, e.g.,* Cullison, *Probability Analysis of Judicial Fact-Finding: A Preliminary Outline of The Subjective Approach,* 1969 U. TOL. L. REV. 538 (1969) [hereinafter cited as Cullison]; Finkelstein & Fairley, *A Bayesian Approach to Identification Evidence,* 83 HARV. L. REV. 489 (1970) [[[hereinafter cited as Finkelstein & Fairley]. *See also,* Becker, *Crime and Punishment: An Economic Approach,* 76 J. POL. ECON. 169 (1968) [hereinafter cited as Becker]; Birmingham, *A Model of Criminal Process: Game Theory and Law,* 56 CORNELL L. REV. 57 (1970) [hereinafter cited as Birmingham]; Kaplan, *Decision Theory and the Fact-finding Process,* 20 STAN. L. REV. 1065 (1968) [[[hereinafter cited as Kaplan]; *cf.* Broun & Kelly, *Playing the Percentages and the Law of Evidence,* 1970 Ill. L. F. 23 [hereinafter cited as Broun & Kelly].

[6] *See, e.g.,* W. WILLS, AN ESSAY ON THE PRINCIPLES OF CIRCUMSTANTIAL EVIDENCE 6-10, 15, 282 (4th ed. 1862); M. HOUTS, FROM EVIDENCE TO PROOF 132 (1956).

[7] *See* the trial testimony of Jan. 18, 1899, and Feb. 4, 1899, reported in a special supplement to Le Petit Temps (Paris), April 22, 1899.

[8] For example, one witness stressed the presence of four coincidences out of the 26 initial and final letters of the 13 repeated polysyllabic words in the document. He evaluated at .2 the probability of an isolated coincidence and calculated a probability of $(0.2)^4$ = .0016 that four such coincidences would occur in normal writing. But $(0.2)^4$ is the probability of four coincidences out of four; that of four or more out of 13 is some 400 times greater, or approximately .7. *See Rappord de Mm. Les Experts Darboux, Appell, et Poincaré,* in LES DOCUMENTS JUDICIARES DE L'AFFAIRE DREYFUS, in LA RÉVISION DU PROCÈS DE RENNES (1909) [hereinafter cited as *Rappord]. Cf.* note 40 *infra.*

[9] Two witnesses observed that, when the word chain "*intérêt/intérêt/intérêt/intérêt/* ...." was compared with the document itself, allowing one letter of slipping-back for each space between words and aligning the word chain with the actual or the ideal left-hand margin as convenient, the letter *l* appeared with particular frequency over the word-chain letter *i;* the letters *n* and *p* appeared frequently over the word-chain letter *n;* and so on. Far from being in any way remarkable, however, the probability that *some* such pattern can be discerned in any document is nearly certainty. *See Rappord* 534.

[10] *See id.*

[11] *See* the discussion of the "selection effect," note 40 *infra.*

[12] A. CHARPENTIER, THE DREYFUS CASE 52-53 (J. May transl. 1935).

[13] *Id.* at 53. *See also id.* at 265.

[14] People v. Collins, 68 Cal. 2d 319, 320, 438 P.2d 33, 66 Cal. Rptr. 497 (1968).

[15] There was testimony that the female defendant's hair color at the time of the robbery was light blond rather than dark blond, as it appeared at trial. The male defendant had no beard at trial or when arrested and told the arresting offcers that he had not worn one on the day of the robbery. There was testimony corroborating his claim that he had shaved his beard approximately two weeks before the robbery, but other testimony that he was bearded the day after the robbery.

[16] The neighbor admitted at trial "that at the preliminary hearing he [[[had] testified to an uncertain identification at the police lineup shortly after the attack ...." 68 Cal. 2d at 321, 438 P.2d at 34, 66 Cal. Rptr. at 498.

[17] *See* explanation in note 63 *infra.*

[18]

Characteristic	***Assumed Probability of its Occurrence***
1. Partly yellow automobile	1/10
2. Man with mustache	1/4
3. Girl with ponytail	1/10
4. Girl with blond hair	1/3
5. Negro man with beard	1/10
6. Interracial couple in car	1/1000

[19] *See* State v. Sneed, 76 N.M. 349, 414 P.2d 858 (1966); People v. Risley, 214 N.Y. 75, 108 N.E. 200 (1915), *discussed at* pp. 1344-45 & notes 47-49 *infra. See also* Campbell v. Board of Educ., 310 F. Supp. 94, 105 (E.D.N.Y. 1970).

[20] The sixth factor, for example, essentially restates parts of the first five. *See* note 18 *supra.*

[21] Precisely this mistake is made in C. MCCORMICK, HANDBOOK OF THE LAW OF EVIDENCE § 171 (1954) and in J. WIGMORE, THE SCIENCE OF JUDICIAL PROOF § 154, at 270-71 (3d ed. 1937). One court has treated such dependence, I think mistakenly, as going only to the "weight" of the product and not to its admissibility. State v. Coolidge, 109 N.H. 403, 419, 260 A.2d 547, 559 (1969), *cert. granted on other issues,* 399 U.S. 926 (1970) (No. 1318 Misc., 1969 Term; renumbered No. 323, 1970 Term), *discussed at* note 40 *infra.*

[22] 68 Cal. 2d at 330, 438 P.2d at 40, 66 Cal. Rptr. at 504.

[23] In a separate mathematical appendix, the court demonstrated that, even if the number of suspect couples approaches only twelve million, the probability that *at least one other couple* (in addition to the actually guilty couple) will possess the six characteristics rises to somewhat over *forty-one percent,* even on the assumption that the prosecutor was correct in concluding that the probability that a randomly chosen couple would possess all six characteristics is but one in twelve million. More generally, the court showed that the probability of such duplication equals

$1-(1-Pr)^{N}-NPr(1-Pr)^{N-1}/1-(1-Pr)^{N}$,

where Pr equals the probability that a random couple will possess the characteristics in question and N is the number of couples in the suspect population. 68 Cal. 2d at 333-35, 438 P.2d at 42-43, 66 Cal. Rptr. at 506-07. If # is taken to represent the value of N • Pr, then the Poisson approximation for the above quotient is $1-\#/e^{\#}-1$, where e is the transcendental number 2.71828.... that is used as the base for natural logarithms. *See* 1 W. FELLER, AN INTRODUCTION TO PROBABILITY THEORY AND ITS APPLICATIONS 153-64 (3d ed. 1968); Kingston, *Applications of Probability Theory in Criminalistics,* 60 J. AM. STATIST. ASS'N 70, 74 (1965). On the assumption that Pr=1/N (so that #=1), the value of the above quotient as N grows without limit is thus (e-2)/(e-1), which is approximately 42, as the court correctly concluded. *See* Cullison, *Identification by Probabilities and Trial by Arithmetic (A Lesson For Beginners in How to be Wrong With Greater Precision),* 6 HOUST. L. REV. 471, 484-502 (1969).

Finkelstein and Fairley suggest that the court's argument was mathematically incorrect because "the court's assumption that one in twelve million is a fair estimate of the probability of selecting such a couple at random necessarily implies that it is a fair estimate of the number of such couples in the population." Finkelstein & Fairley 493; *accord,* Broun & Kelly, *supra* note 5, at 43. But this completely misconceives the argument. Of course, if the figure of one in twelve million had represented an estimate, based upon random sampling, of *the actual frequency of Collins-like couples in a known population,* the criticism of the court's opinion would be well taken. But in fact the "one-in-twelve-million" figure represented nothing of the sort. Since nothing was known about exactly who was and who was not a member of the population of "suspect" couples, that figure represented only an estimate of the probability that any given couple, chosen at random from an unknown population of "suspect" couples, would turn out to have the six "Collins" characteristics--with that estimate itself based only on a multiplication of component factors, each representing the frequency of one of the six characteristics in a much larger population.

[24] *See also* note 40 *infra.*

[25] The court stressed the fact that the prosecutor had criticized the traditional notion of proof beyond a reasonable doubt as "hackneyed" and "trite"; that he "sought to reconcile the jury to the risk that, under his 'new math' approach to criminal jurisprudence, 'on some rare occasion … an innocent person may be convicted'"; and that he thereby sought "to persuade the jury to convict [the] defendants whether or not they were convinced of their guilt to a moral certainty and beyond a reasonable doubt." 68 Cal. 2d at 331-32, 438 P.2d at 41, 66 Cal. Rptr. at 505. The interaction between mathematical proof and reasonable doubt is discussed at pp. 1372-75 *infra.*

[26] 68 Cal. 2d at 320, 438 P.2d at 33, 66 Cal. Rptr. at 497.

[27] *See, e.g.,* 1967 DUKE L.J. 665, 681-83, *discussing* State v. Sneed, 76 N.M. 349, 414 P.2d 858 (1966).

[28] *E.g.,* United States v. United Shoe Machinery Corp., 110 F. Supp. 295, 304-05 (D. Mass. 1953) (Wyzanski, J.), *aff'd per curiam,* 347 U.S. 521 (1954).

[29] *See, e.g.,* Louisville & N.R.R. v. Steel, 257 Ala. 474, 59 So. 2d 664 (1952); Von Tersch v. Ahrendsen, 251 Iowa 115, 99 N.W.2d 287 (1959).

[30] *See, e.g.,* United States v. 88 Cases, More or Less, Containing Bireley's Orange Beverage, 187 F.2d 967, 974 (3d Cir. 1951).

[31] *See generally,* Finkelstein, *The Application of Statistical Decision Theory to the Jury Discrimination Cases,* 80 HARV. L. REV. 338 (1966); Zeisel, *Dr. Spock and the Case of the Vanishing Women Jurors,* 37 U. CHI. L. REV. 1 (1969) [hereinafter cited as Zeisel]. *But see* State v. Smith, 102 N.J. Super. 325, 341, 246 A.2d 35, 50 (1968), *aff'd,* 55 N.J. 476, 262 A.2d 868 (1970), *cert. denied,* 400 U.S. 949 (1970).

[32] *See* pp. 1365-67, p. 1381 & notes 33, 37 & 41 *infra.*

[33] A sensible, and now quite conventional, approach to this question is "to treat the probability as the fact if the defendant has the power to rebut the inference." Jaffe, *Res Ipsa Loquitur Vindicated,* 1 BUFF. L. REV. 1, 6 (1951). On this theory, if the defendant produces a reasonably satisfactory explanation consistent with a conclusion of no negligence, and if the plaintiff produces no further evidence, the plaintiff should lose on a directed verdict despite his mathematical proof--unless (1) he can adequately explain his inability to make a more particularized showing (a possibility not adverted to in *id.*), or (2) no specific explanation is given, but there is some policy reason to ground liability in the area in question on a substantial probability of negligence in the *type* of case rather than to require a reasoned probability in the *particular* case, *cf.* note 100 *infra,* thereby moving toward a broader basis of liability. It will be noticed that no such policy is likely to operate when the mathematical evidence goes to the question of the defendant's *identity* and the plaintiff does not explain his failure to produce any more particularized evidence, for it will almost always be important to impose liability on the correct party, *whatever* the basis of such liability might be. *See* p. 1349 *infra. See also* notes 37 & 102 *infra.*

[34] It has now been settled as a federal constitutional matter, *see* Turner v. United States, 396 U.S. 398 (1970), that this statistical fact permits a legislature to authorize a jury to find illegal importation once it finds possession "unless the defendant explains the possession to the satisfaction of the jury." 21 U.S.C. § 174 (1964); *cf.* Leary v. United States, 395 U.S. 6 (1969). At least one

commentator has urged the alternative position that the jury should not in such cases be instructed that proof of possession is sufficient to find illegal importation (for that shifts to the accused the practical burden of persuasion, with its accompanying pressure to testify, notwithstanding any contrary jury charge) but should instead be told that ninety-eight percent of all heroin in the United States is illegally imported (for that leaves the jury more likely to give even the non-testifying accused the benefit of the doubt created by the remaining two percent). Comment, *Statutory Criminal Presumptions: Reconciling the Practical With the Sacrosanct,* 18 U.C.L.A. L. REV. 157 (1970). But it is by no means clear, despite the commentator's assertion, that "the jury is more likely to consider other relevant circumstances unique to the particular case on a more equal footing with the 98 percent statistic than it would with a presumption." *Id.* at 183 n.102. *See generally* the discussion at pp. 1359-65 *infra.*

[35] If tires rotated in complete synchrony with one another, the probability would be 1/12; if independently, 1/12 x 1/12, or 1/144.

[36] A Swedish court, computing the probability at 1/12 x 1/12 = 1/144 on the dubious assumption that car wheels rotate independently, ruled that fraction large enough to establish reasonable doubt. Parkeringsfrägor, II. Tilförlitligheten av det S.K. locksystemet för parkernigskontroll. *Svensk juristidining,* 47 (1962) 17-32, *cited in* Zeisel, *supra* note 31, at 12. The court's mathematical knife cut both ways, however, for it added that, had all four tire-valves been recorded and found in the same position, the probability of 1/12 x 1/12 x 1/12 x 1/12 = 1/20, 736 would have constituted proof beyond a reasonable doubt. *Id.* For a discussion of why no such translation of the "reasonable doubt" concept into mathematical terms should be attempted, see pp. 1372-75 *infra.*

[37] In Smith v. Rapid Transit, Inc., 317 Mass. 469, 58 N.E.2d 754 (1945), the actual case on which this famous chestnut is based, no statistical data were in fact presented, but the plaintiff did introduce evidence sufficient to show that the defendant's bus line was the only one chartered to operate on the street where the accident occurred. Affirming the direction of a verdict for the defendant, the court observed: "The most that can be said of the evidence in the instant case is that perhaps the mathematical chances somewhat favor the proposition that a bus of the defendant caused the accident. This was not enough." 317 Mass. at 470, 58 N.E.2d at 755. *See also* Sawyer v. United States, 148 F. Supp. 877 (M.D. Ga. 1956); Reid v. San Pedro, L.A.& S.L.R.R., 39 Utah 617, 118 P. 1009 (1911). If understood as insisting on a numerically higher showing--an "extra margin" of probability above, say, .55-- then the decision in *Smith* would make no sense, at least if the court's objective were the minimization of the total number of judicial errors in situations of this kind, an objective essentially implicit in the adoption of a "preponderance of the evidence" standard. *See* Ball, *The Moment of Truth: Probability Theory and Standards of Proof,* 14 VAND. L. REV. 807, 822-23 (1961) [[[[hereinafter cited as Ball]. But cases like *Smith* are entirely sensible if understood instead as insisting on the presentation of *some* non-statistical and "individualized" proof of identity before compelling a party to pay damages, and even before compelling him to come forward with defensive evidence, absent an adequate explanation of the failure to present such individualized proof. *Compare* p. 1349 *infra with* note 33 *supra.*

[38] Note that in this criminal case, as in the preceding civil one, *a fact known about the particular defendant* provides reason to believe that the defendant is involved in a certain percentage of all cases (here, cases of being at the crucial place between 7 p.m. and midnight) possessing a characteristic shared by the litigated case.

[39] In this case, unlike the preceding two, it is *a fact known about the particular event that underlies the litigation,* not any fact known about the defendant, that triggers the probabilistic showing: a certain percentage of all events in which the crucial fact (here, the killing of a man in his mistress' apartment) is true are supposedly caused by a person with a characteristic (here, being the mistress) shared by the defendant in this case.

[40] This is, of course, People v. Collins, 68 Cal. 2d 319, 438 P.2d 33, 66 Cal. Rptr. 497 (1968), minus the specific mathematical errors of *Collins* and without the interracial couple. One special factor that can lead to major mathematical distortions in this type of case is the "selection effect" that may arise from either party's power to choose matching features for quantification while ignoring non-matching features, thereby producing a grossly exaggerated estimate of the improbability that the observed matching would have occurred by chance. *See* Finkelstein & Fairley 495 n.14. This difficulty may well have been present in People v. Trujillo, 32 Cal. 2d 105, 194 P.2d 681, *cert. denied,* 335 U.S. 887 (1948), in which an expert examined a large number of fibers taken from clothing worn by the accused and concluded, upon finding eleven matches with fibers taken from the scene of the crime, that there was only a one-in-a-billion probability of such matching occurring by chance. A particularly egregious case of this sort is State v. Coolidge, 109 N.H. 403, 260 A.2d 547 (1969), *cert. granted on other issues,* 399 U.S. 926

(1970) (No. 1318 Misc., 1969 Term; renumbered No. 323, 1970 Term), where particles taken from the victim's clothing were found to match particles taken from the defendant's car and clothing in twenty-seven out of forty cases.

In expressing his conclusion based upon statistical probabilities, the [[[[consultant in micro-analysis and director of a university laboratory for scientific investigation] relied upon previous studies made by him, indicating that the probability of finding similar particles in sweepings from a series of automobiles was one in ten. Applying this as a standard, he determined the probability of finding 27 similar particles in sweepings from independent sources would be only one in ten to the 27th power.

109 N.H. at 418-19, 260 A.2d at 559. The court upheld the admissibility of that testimony, 109 N.H. at 422, 260 A.2d at 561, notwithstanding the weakness of the underlying figure of 1/10 and the expert's own concession that the particle sweepings "may not have been wholly independent," 109 N.H. at 419, 260 A.2d at 559. *See* note 21 *supra.* Most significantly, the court was evidently unaware that the relevant probability, that of finding 27 or more matches *out of 40 attempts,* was very much larger than $1/10^{27}$--larger, in fact, by a factor of approximately 10^{10}. Indeed, even the 40 particles chosen for comparison were visually selected for similarity from a still larger set of particle candidates, 109 N.H. at 421, 260 A.2d at 560--so large a set, conceivably, that the probability of finding 27 or more matches in sweeping over such a large sample, even from two entirely different sources, could well have been as high as 1/2 or more. *Cf.* note 8 *supra.* Oddly, the expert testimony in *Coolidge* has recently been described as "not misleading." Broun & Kelly, *supra* note 5, at 48.

[41] It is, of course, a fair question how such evidence could ever be compiled; the difficulty, and perhaps the impossibility, of compiling it no doubt reflects the "nonobjective" nature of the intent inquiry. *See* pp. 1365-66 *infra.*

[42] *See* p. 1339 *supra.*

[43] Turner v. United States, 396 U.S. 398 (1970), sustained an authorized jury inference of knowledge in these circumstances. *See* note 34 *supra.*

[44] *See* note 32 *supra.*

[45] *See, e.g.,* People v. Trujillo, 32 Cal. 2d 105, 194 P.2d 681, *cert. denied,* 335 U.S. 887 (1948); State v. Coolidge, 109 N.H. 403, 260 A.2d 547 (1969), *cert. granted on other issues,* 399 U.S. 926 (1970) (No. 1318 Misc., 1969 Term; Renumbered No. 323, 1970 Term), *discussed in* note 40 *supra. See also* Note, *The Howland Will Case,* 4 AM. L. REV. 625 (1870), *discussing* Robinson v. Mandell, 20 F. Cas. 1027 (No. 11,959) (C.C.D. Mass. 1868), *discussed in* note 47 *infra;* People v. Jordan, 45 Cal. 2d 697, 707, 290 P.2d 484, 490 (1955), *discussed in* note 155 *infra.*

[46] *See, e.g.,* People v. Collins, 68 Cal. 2d 319, 438 P.2d 33, 66 Cal. Rptr. 497 (1968), *discussed at* pp. 1334-37 *supra;* State v. Sneed, 76 N.M. 349, 414 P.2d 858 (1966); People v. Risley, 214 N.Y. 75, 108 N.E. 200 (1915). *See also* Smith v. Rapid Transit, Inc., 317 Mass. 469, 58 N.E.2d 754 (1945), *discussed in* note 37 *supra;* Miller v. State, 240 Ark. 340, 399 S.W.2d 268 (1966), *discussed in* note 155 *infra.*

[47] People v. Risley, 214 N.Y. 75, 108 N.E. 200 (1915). Experts on typewriters had been called to testify that certain peculiarities in the forged document corresponded completely with peculiarities in a typed sample produced by the defendant's typewriter. *Cf.* State v. Freshwater, 30 Utah 442, 85 P. 447 (1906). A mathematician was then allowed to testify, in response to a hypothetical question ascribing certain probabilities to the occurrence of any one defect in a random typewriter, that the probability of the coincidence of all these defects in any single machine was but one in four billion. *Cf.* Note, *supra* note 45, at 648-49 (1870), *discussing* Robinson v. Mandell, 20 F. Cas. 1027 (No. 11,959) (C.C.D. Mass 1868), in which Benjamin Pierce, Harvard Professor of Mathematics, applied the product rule to strokes of authentic and disputed signatures to conclude that their similarities should be expected to occur by chance only once in a number of times equal to the thirtieth power of five. The *Risley* court found reversible error in the use of the "one-in-four-billion" argument on the narrow and obviously correct ground that the hypothetically assumed probabilities for the separate defects were unsupported by any evidence in the case. But the court went on to indicate its view, quoted in text above, that probabilistic evidence is necessarily inadmissible to establish a past event.

[48] 214 N.Y. at 86, 108 N.E. at 203. If the court's use of the phrase "should still be alive" is taken to suggest that *A*'s death has otherwise been firmly established, then the court's example has a surface plausibility arising not out of the fact that *A*'s alleged death is past rather than future, but out of the obviously low probative force of general statistical averages of this sort when confronted with convincing evidence more narrowly focused on the disputed event itself. *Cf.* note 100 *infra.* If, on the other hand, the court's assertion is taken to deny the relevance of life expectancy data in deciding a genuine factual dispute as to whether or not *A* had died, then the court's denial flies in the face of at least the theory that underlies the traditional presumption of death in cases of long and unexplained absence. *See* Comment, *A Review of the Presumption of Death in New York,* 26 ALBANY L. REV. 231, 245 (1962). *See also* note 101 *infra.*

[49] 214 N.Y. at 86, 108 N.E. at 203. *But see* Liddle, *Mathematical and Statistical Probability As a Test of Circumstantial Evidence,* 19 CASE W. RES. L. REV. 254, 277-78 (1968), expressing the surprising view that "[m] athematical probability is … most useful in establishing the existence of or identifying facts relating to past events and least useful in the predicting of future events …."

[50] *See* note 47 *supra.*

[51] *See* Note, *Evidential Use of Mathematically Determined Probability,* 28 HARV. L. REV. 693, 695-96 (1915).

[52] Ball, *supra* note 37, at 815 n.19, *citing* A. AMES, THE MORNING NOTES OF ADELBERT AMES, JR. (H. Cantril ed. 1960).

[53] Ball, *supra* note 37, at 815.

[54] *Cf.* note 100 *infra* for a related but more supportable proposition.

[55] If no other evidence has been adduced by the time the plaintiff has rested, we at least know the very fact that no other evidence has been adduced--a fact that may properly be treated as dispositive in some situations if other evidence on the issue of identity seems likely to have been available. *Cf.* Case v. New York Central R.R., 329 F.2d 936, 938 (2d Cir. 1964) (Friendly, J.); National Life & Accident Ins. Co. v. Eddings, 188 Tenn. 512, 221 S.W.2d 695 (1949). *See* p. 1349 *infra.*

[56] *See* note 33 *supra* and note 102 *infra.*

[57] L. SAVAGE, FOUNDATIONS OF STATISTICS (1950).

[58] The notion of probabilities as measures of an individual's "degree of confidence" or "degree of belief" in an uncertain proposition or event traces to JAMES BERNOULLI, ARS CONJECTANDI (1713). See the extremely helpful historical discussion in H. RAIFFA, DECISION ANALYSIS, INTRODUCTORY LECTURES ON CHOICES UNDER UNCERTAINTY 273-78 (1968) [hereinafter cited as RAIFFA]. The special contribution of Savage was to formalize that notion in terms useful for the rigorous study of decisionmaking under uncertainty.

[59] It is possible to know *only* this fact at the outset of the trial-- though not, of course, at its conclusion, *see* note 55 *supra*--unless the bulk of the trial record were somehow destroyed and, with it, all memory of what had been established during the proceedings.

[60] To make the idea of a "bet" involving B_{80} correspond as nearly as possible to the situation that actually confronts the trier of fact, one need only postulate that the reward accorded a correct guess consists in learning that a particular lawsuit which the trier wants to see rightly decided has in fact been correctly determined.

[61] Equating the defendant's liability with the mere fact of identity may, of course, overlook other important elements of legal responsibility, a matter taken up at pp. 1365-66 *infra.* I make the equation here only on the explicit assumption that identity is the sole issue in the litigation.

[62] Typical is the "transitivity" postulate that one who regards *X* as more probable than *Y,* and *Y* as more probable than *Z,* should also regard *X* as more probable than *Z.* The other postulates simply specify that P(X) is never less than zero or greater than one and that, if *A* and *B* are any two mutually exclusive propositions, P(A) + P(B) equals P(A or B).

[63] Perhaps the most involved of the few basic rules that make the translation a useful one is the rule used to calculate the probability that two propositions, *A* and *B,* are both true. If a person learned somehow that *A* was in fact true, how sure would he *then* be of *B*'s truth? The answer to that question, calibrated in terms of the sequence of boxes described above, measures the probability, for this person, of *B conditioned on A,* or of *B given A,* written P(B|A). *See* note 71 *infra.* The rule in question simply states that the person's probability estimate of the *joint* truth of *A* and *B* equals his probability estimate of *A,* multiplied by his probability estimate of *B* conditioned on *A.* Symbolically, P(A&B) = P(A) • P(B|A). For example, if a person thinks that *A* is exactly as likely to be true as false (*i.e.,* P(A) = 1/2), and if learning of *A*' s truth would lead him to think that *B* is only half as likely to be true as false (*i.e.,* P(B|A) = 1/3), then he should conclude that the probability of *A* and *B both* being true is (1/2) • (1/3) = 1/6. In the special case where he believes *A* and *B* are *mutually independent,* his knowledge of *A*'s truth would have no effect on his estimate of *B,* so that, for him, P(B|A) would equal P(B), and thus P(A&B) would simply equal P(A) • P(B)--which is the "product rule," *noted at* p. 1335 *supra.*

It should be noted here that Finkelstein and Fairley suggest that no complete translation from objective frequencies to subjective probabilities is actually re-required, for they theorize that even subjective probabilities may be interpreted as expressing frequencies. Thus, they equate the statement that the subjective probability of the defendant's guilt is 1/2 with the statement that "if a jury convicted whenever the evidence generated a similar degree of belief in guilt, the verdicts in this group of cases would tend to be right about half the time." Finkelstein & Fairley 504. *See also* Broun & Kelly, *supra* note 5, at 31; Kaplan 1073. But the functional relationship between subjective probabilities and likely outcomes is far more complex than this equation assumes, for it turns on such factors as how easy or difficult it is for either party to generate a given level of belief in a false propositon. *See* the related discussion at p. 1385 *infra.*

[64] *See* p. 1340 *supra.*

[65] *See* note 37 *supra.* Indeed, some statistical evidence, *see, e.g.,* note 48 *supra,* is so general and remote from the particular case as to be of only marginal relevance--if that. Of course, if plaintiff can satisfactorily account for the evidentiary omission, the statistical evidence alone might well suffice. *See* note 102 *infra.* And what constitutes a satisfactory explanation of the failure to adduce non-statistical evidence might *itself* turn, at least in part, on the level of probability suggested by the statistics. This possibility was not noted by Ball, *supra* note 37, at 822-823.

[66] *See id.* at 823. This seems to me the only sensible meaning that can be attached to such pronouncements as that "a verdict must be based upon what the jury finds to be facts rather than what they find to be more 'probable'." Lampe v. Franklin Am. Trust Co., 339 Mo. 361, 384, 96 S.W.2d 710, 723 (1936). *Accord,* Frazier v. Frazier, 228 S.C. 149, 168, 89 S.E.2d 225, 235 (1955). *See also* Note, *Variable Verbalistics--the Measure of Persuasion in Tennessee,* 11 VAND. L. REV. 1413 (1958).

[67] There are, of course, possibilities of proportioned verdicts or other forms of judicial compromise in obviously doubtful cases, but these are not without their own subtle difficulties. *See* Allen, Coons, Freund, Fuller, Jones, Kaufman, Nathanson, Noonan, Ruder, Schuyler, Sowle & Snyder, *On Approaches to Court Imposed Compromises--The Uses of Doubt and Reason,* 58 NW. U.L. REV. 750, 795 (1964).

[68] People v. Collins, 68 Cal. 2d 319, 330, 438 P.2d 33, 40, 66 Cal. Rptr. 497, 504 (1968).

[69] Reverend Thomas Bayes, in *An Essay Toward Solving a Problem in the Doctrine of Chance,* PHILOSOPHICAL TRANS. OF THE ROYAL SOCIETY (1763), suggested that probability judgments based on intuitive guesses should be combined with probabilities based on frequencies by the use of what has come to be known as Bayes' Theorem, a fairly simple formula that is derived at p. 1352 *infra.* More recently, it has become common to think of Bayes' Theorem as providing "a quantitative description of the ordinary process of weighing evidence." I. GOOD, PROBABILITY AND THE WEIGHING OF EVIDENCE 62 (1950). *See also* J. VENN, THE LOGIC OF CHANCE ch. 16-17 (3rd ed. 1868).

[70] P(X|E) is usually read "*P* of *X* given *E,*" or "the probability of *X* given the truth of *E.*"

[71] A conditional probability like $P(X|E)$ is often understood to assume the given condition *E* as a certainty. One can as readily interpret $P(X|E)$, however, as measuring the degree to which the trier would believe *X if* he were sure of *E. See* note 63 *supra.*

[72] To make these formulas intuitively transparent, consider exactly what they assert. The first asserts that the probability that *A* and *B* are *both* true equals the product of two other probabilities: the probability that *B* is true, multiplied by the probability that *A* would also be true *if B* were true. *See* notes 63 & 71 *supra.*

The second formula asserts that the probability that *A* is true equals the sum of two other probabilities: the probability that *A* is true and *B* is *true,* plus the probability that *A* is true and *B* is *false.* Of course, *B* is either true or false, mutually exclusive possibilities, so the second formula reduces to the assertion that the probability that one of two mutually exclusive events will occur equals the sum of the probabilities of each event's occurrence. For example, if a well-shuffled deck contains ten white cards, ten gray cards, and eighty black cards, the probability that a card chosen randomly from the deck will be either white or gray equals .2, the sum of the probability that it will be white (.1) and the probability that it will be gray (.1).

[73] Formula (1) implies that $P(E\&X) = P(E|X) \bullet P(X)$. But E&X is identical with X&E so $P(E\&X) = P(X\&E) = P(X|E) \bullet P(E)$. Thus $P(X|E) \bullet P(E) = P(E|X) \bullet P(X)$, from which we obtain formula (3) by dividing P(E) into both sides of this equation. Formula (2) implies that $P(E) = P(E\&X) + P(E \,\&\, \text{not-}X)$. Applying formula (1), we know that $P(E\&X) = P(E|X) \bullet P(X)$ and that $P(E \,\&\, \text{not-}X) = P(E|\text{not-}X) \bullet P(\text{not-}X)$, from which we obtain formula (4) by adding these two terms.

[74] The only other variable in (5), P(not-X), is equal to 1-P(X).

[75] Another formulation of Bayes' Theorem, less conventional and for most purposes less convenient, might nonetheless be noted here inasmuch as it may be easier to grasp intuitively. If O(X) represents the "odds of *X,*" defined as P(X)/P(not-X), then

$O(X|E) = P(E|X)/P(E|\text{not-}X) \bullet O(X)$,

which in effect defines the "probative force" of *E* with respect to *X,* written PF(E wrt X), as the ratio of the probability that *E* would be true if *X* were *true* to the probability that *E* would be true if *X* were *false.* It is tempting, but incorrect, to assume that $PF(E_1 \,\&\, E_2 \text{ wrt } X)$, the probative force of the combination of E_1 and E_2 with respect to *X,* always equals the product of their separate probative forces, $PF(E_1 \text{ wrt } X) \bullet PF(E_2 \text{ wrt } X)$. If E_1 and E_2 are not *conditionally independent* of both *X* and not-X (*i.e.,* if $P(E_1 \,\&\, E_2|X) \neq P(E_1|X) \bullet P(E_2|X)$ *or* if $P(E_1 \,\&\, E_2|\text{not-}X) \neq P(E_1|\text{not-}X) \bullet P(E_2|\text{not-}X)$), then one *cannot conclude* that

$O(X|E_1 \,\&\, E_2) = PF(E_1 \text{ wrt } X) \bullet PF(E_2 \text{ wrt } X) \bullet O(X)$,

although the above formula *does* hold if conditional independence obtains (*i.e.,* if $P(E_1 \,\&\, E_2|X) = P(E_1|X) \bullet P(E_2|X)$ *and* $P(E_1 \,\&\, E_2|\text{not-}X) = P(E_1|\text{not-}X) \bullet P(E_2|\text{not-}X)$). *See* pp. 1366-68 *infra.* These difficulties are overlooked in Kaplan 1085-86.

[76] Finkelstein & Fairley, *supra* note 5.

[77] *Id.* 502, 516-17. Although he is somewhat less precise about his suggestion, another commentator may have intended to advance a similar proposal in an earlier article. *See* Cullison, *supra* note 23, at 505. And a roughly equivalent proposal was in fact put forth over twenty years ago. *See* I. GOOD, PROBABILITY AND THE WEIGHING OF EVIDENCE 66-67 (1950).

[78] 68 Cal. 2d at 320, 438 P.2d at 33, 66 Cal, Rptr. at 497.

[79] Finkelstein & Fairley 497.

[80] Contrary to the implication in *id.,* nothing whatever in the *Collins* opinion suggests that the palm-print evidence *itself* would be thought to have insufficient probative value to be admissible. Had that been the view of the *Collins* court, it would have been forced to conclude that the evidence of the six matching characteristics was inadmissible on the facts of that case. Of course, the court concluded no such thing, and rejected only the prosecutor's particular attempt to *quantify* the probative force of the coincidence of characteristics. *Cf.* note 2 *supra.*

[81] *See* pp. 1336-38 *supra.*

[82] In fairness, it should be said that the solution is put forth quite tentatively, *see* Finkelstein & Fairley 502, and that, although Finkelstein and Fairley do expressly advocate its adoption, *see id.* 516-17, the main thrust of their work is to enlarge our understanding of the process of evidentiary inference through the application of Bayesian techniques, a task that they perform

admirably and with a candor and explicitness that makes it possible for others to criticize and build upon their initial efforts. My disagreement is only with their suggested use of those techniques, however improved, *at trial.*

[83] *Id.* 498.

[84] *Id.* 500.

[85] *See* pp. 1362-65 *infra.*

[86] I do not argue that the methods in question will invariably display any intrinsic bias *outside* the trial context, at least insofar as their use can be subjected to continuing scrutiny and improvement over time. In the trial setting, however, there are institutional goals and constraints that effectively preclude the undistorted use of mathematical techniques. *See generally* pp. 1358-77 *infra.*

[87] Finkelstein & Fairley 500.

[88] *Id.* 502. If the print's frequency were disputed, the jurors would of course have to be shown more than one row of such a table.

[89] *Id.* 502-03 n.33.

[90] *Id.* 517.

[91] *See* p. 1346 *supra.*

[92] *See* Finkelstein & Fairley 498-505. The authors do assert that "[u] nder certain restricted conditions, useful prior probabilities can be estimated on the basis of objective population statistics without resort to subjective evaluations," *id.* 506, and discuss studies of the use of statistics to determine prior probabilities in Polish paternity suits. *Id.* 506-09. But while discussing others' suggestions of the uses judges could make of such statistically based prior probability assessments, Finkelstein and Fairley do not consider the application of their technique of using Bayesian analysis at trial to such prior probabilities, nor consider its implications. Moreover, the authors suggest that, even if statistics could be used to arrive at an objective prior probability, "[w]here in [the judge's] opinion the facts showed that the case was either stronger or weaker than usual, he could subjectively adjust the prior [probability] accordingly." *Id.* 509.

[93] *E.g.,* the palm-print hypothetical, *discussed at* pp. 1355-58 *supra,* or the *Collins* case, *discussed at* pp. 1334-37 *supra,* or the Thunderbird hypothetical, *discussed at* p. 1342 *supra.*

[94] *E.g.,* the parking hypothetical, *discussed at* p. 1340 *supra.*

[95] Finkelstein & Fairley 517.

[96] *See* p. 1339 *supra.*

[97] *See* pp. 1339-40 & p. 1343 *supra.*

[98] *See* pp. 1340-41 *supra.*

[99] *See* p. 1341 *supra.*

[100] When the mathematical evidence *E* simply describes the *class* of cases of which the litigated case is one, the truth or falsity of the litigated proposition *X* in no way affects the probability of the truth of *E; i.e.,* P(E|X) equals P(E). Bayes' Theorem then simply asserts that P(X|E) = P(X). *See* equation (3), p. 1352 *supra.* Hence, in all such cases, *E* can suggest what *initial* value to assign to P(X) but cannot serve to *refine* that initial value-- as can, for example, evidence of the palm-print variety. This formulation makes more precise the common-sense notion, *cf.* p. 1346 *supra,* that the sort of statistical evidence that was offered in the bus case pertains not to the *particular* dispute being litigated but to a broad *category* of possible disputes. Evidence *E* is of this character if, although *E* is *relevant* to the disputed proposition *X,* it is nonetheless true that P(E|X) = P(E).

[101] For example, in line with the suggested approach, the judge might decide to employ the doctrine of *res ipsa loquitur, see* note 33 *supra,* or any of a variety of rebuttable presumptions. *See, e.g.,* O'Dea v. Amodeo, 118 Conn. 58, 170 A. 486 (1934) (presumption of father's consent to son's operation of automobile); Hinds v. John Hancock Mut. Life Ins. Co., 155 Me. 349, 354-67, 155 A.2d 721, 725-32 (1959) (presumption against suicide). One of the traditional functions of the use of presumptions, at least those rebuttable by any substantial contrary evidence, is "to make more likely a finding in accord with the balance of probability." Morgan, *Instructing the Jury upon Presumptions and Burden of Proof,* 47 HARV. L. REV. 59, 77 (1933).

[102] *See* p. 1349 *supra.* If the statistical evidence standing alone establishes a sufficiently high prior probability of *X,* and a satisfactory explanation is provided for the failure to adduce more individualized proof, there seems no defensible alternative (absent believable evidence contrary to *X*) to directing a verdict for the party claiming *X,* for no factual question remains about which the jury can reason, and directing a verdict the other way would be more likely to lead to an unjust result. If, however, more individualized proof *is* adduced, and if the party opposing *X* has discharged the burden (created by the statistical evidence, *see* note 33 *supra*) of producing believable evidence to the contrary, the question remains whether the risk of distortion created by informing the trier of fact of the potentially overbearing statistics so outweighs the probative value of such statistics as to compel their judicial exclusion. If this situation arises in a criminal case, *see, e.g.,* the heroin hypotheticals, p. 1340 & p. 1343 *supra;* the police hypothetical, p. 1341 *supra;* and the mistress hypothetical, *id.,* the added threats to important values, *see* pp. 1368-75 *infra,* should probably suffice, in combination with the danger of a distorted outcome, to outweigh the probative value of the statistics. But if the situation arises in a civil case, as in the barrel hypothetical, p. 1339 *supra,* or in the bus hypothetical, p. 1340 *supra,* all that I am now prepared to say is that the question of admissibility seems to me a very close one.

[103] Finkelstein & Fairley 498.

[104] *Id.* 509-11.

[105] The same palm might leave a variety of seemingly different partial prints.

[106] Significantly, the authors take account of the problem of source variations, Finkelstein & Fairley 510, but neglect to consider the other, less readily quantifiable, components of the variable *g. See id.* 498 n.22.

[107] *Id.* 498, 500.

[108] Even this is somewhat oversimplified since it neglects to multiply the frequency by the factor (1-g). *See also* note 113 *infra.*

[109] *See* p. 1336 *supra.*

[110] The frame-up possibility was also overlooked by Broun & Kelly, *supra* note 5, at 27-28 & n.20. Finkelstein and Fairley seem to have overlooked this possibility literally by definition. They define *X* (the authors labeled it *G; see* pp. 1365-66 *infra*) to be "the event that defendant used the knife," apparently meaning that the defendant used the knife to kill. Finkelstein & Fairley 498. Yet they define not-X to be the event "that a palm print [was] left by someone other than the defendant," *id.,* leaving the frame-up case as one where both *X* and not-X are untrue, a logical impossibility.
Of course, it is possible that the authors intentionally prevented the frame-up possibility from affecting their calculations by the use of the definition of *X,* "the event that defendant used the knife," to mean the event that the print was the defendant's--in which case the frame-up case is included within *X.* But this would require another application of Bayes' Theorem to determine the probability of guilt from the knowledge of the probability that the print was the defendant's, a step that Finkelstein and Fairley clearly did not intend. Instead, they repeatedly referred to P(X) as the "prior probability of guilt." *See, e.g., id.* 500; *see* pp. 1365-66 *infra.*

[111] A quite distinct problem, more serious in cases relying on human identification than in cases where identification is based on physical evidence, is that the characteristics of the defendant relied upon in the probabilistic formula may not in fact have been shared by the actually guilty individual or individuals because of some mistake in observation or memory. Such risks of error, like the risk of frame-up, are hard to quantify and hence likely to be underemphasized in a quantitative analysis, but differ from the risk of frame-up in that they need not perversely increase as the apparent probative value of the evidence increases.

[112] *Cf.* pp. 1337-38 *supra.* This is not to say, of course, that a professional decisionmaking body, possessed of both the skills and the time to refine its sophistication in mathematical techniques, could not overcome such a tendency--but the jury is not such a body, and making it into one for all but a very limited set of purposes would entail staggeringly high costs.

[113] More precisely, $P(E|\text{not-}X)$ equals not simply $f+F$, but $(1-g)f+F$. *Cf.* note 108 *supra.* This fact makes the final computation even more complex, although it changes the outcome only slightly if g is small in relation to f or F.

[114] *See* equation (5), p. 1352 *supra.* The complete equation, taking into account note 113 *supra,* is more involved still:
$P(X|E) = [(1-g)/(1-g)\ P(X) + [(1-g)\ f + F]\ P(\text{not-}X)] \cdot P(X)$.

[115] Indeed, if $F = .25$, $f = .1$, $g = .1$, and $P(X) = .25$, then $P(X|E) = .469$, not .769 (as Finkelstein and Fairley calculated, *id.* 500)--this time enough of a difference to cross even the “preponderance” line of .5, making $P(X\ |E)$ *less* likely than not rather than *more* likely than not, as the Finkelstein-Fairley method would suggest. As Professor Howard Raiffa pointed out to me upon reading this argument, it highlights how useful mathematics can be in illuminating the source, size, and structure of such distortions and underscores the point that my objection is not to mathematical analysis as such but to its formal use at trial. *See also* note 132 *infra.*

[116] *See* pp. 1375-77 *infra.* Further complicating the picture, there will often be dispute as to the truth of such underlying evidentiary facts as *E* which call forth the use of statistics, a circumstance that makes the use of Bayes' Theorem more complex still.

[117] *Cf.* Lipsey & Lancaster, *The General Theory of Second Best,* 24 REV. ECON. STUDIES 11 (1956) (economic theory that, once some constraint prevents attainment of one optimum condition, other previously optimal conditions are generally no longer desirable as means to best solution):
Specifically, it is *not* true that a situation in which more, but not all, of the optimum conditions are fulfilled is necessarily, or is even likely to be, superior to a situation in which fewer are fulfilled.
Id. 12.

[118] Finkelstein & Fairley 498-500.

[119] Although such matters will rarely be at issue in cases where the defense rests in part on a claim of mistaken identity, it is of course possible for both identity and intent to be disputed in the same case.

[120] *See* pp. 1342-43 *supra.*

[121] This difficulty would be somewhat easier to correct than any of the others identified thus far--through the use of special instructions to the jury, or perhaps through special verdicts.

[122] People v. Collins, 68 Cal. 2d 319, 321, 438 P.2d 33, 34, 66 Cal. Rptr. 497, 498 (1968).

[123] Strictly speaking, the requirement is that *E* be *conditionally* independent of both *X* and not-X, where *X* is the proposition in dispute. *See* note 75 *supra.*

[124] An illustration of the converse situation, in which $P(X|E_1)$ and $P(X|E_2)$ are both smaller than $P(X)$ but $P(X|E_1\ \&\ E_2)$ is much larger, is easy to construct. E_1 and E_2 could, for example, represent mutually inconsistent but nonetheless strikingly similar alibis. Independently, the alibis might seem plausible, thus reducing the probability of *X.* Taken together, not only does the inconsistency destroy the effectiveness of each in reducing the likelihood of *X,* but the similarity makes both seem contrived, allowing a more probable inference of *X.*

[125] By successive applications of equations (1) and (3), pp. 1351-52 *supra:*

$$\mathbf{P(X|E_1\ \&\ E_2) = P(X) \cdot P(E_1\ \&\ E_2|X)/P(E_1\ \&\ E_2) = P(X\ \&\ E_1\ \&\ E_2)/\ P(E_1\ \&\ E_2) = P(X\ \&\ E_1) \cdot P(E_2|X\ \&\ E_1)/\ P(E_1\ \&\ E_2) = P(E_1) \cdot P(X\ |E_1) \cdot P(E_2|X\ \&\ E_1)/P(E_1) \cdot P(E_2|E_1) = P(E_2|X\ \&\ E_1) \cdot P(X|E_1)/\ P(E_2|E_1).}$$

[126] This limitation is simply ignored by Kaplan, *supra* note 5, at 1084-85; *see* note 75 *supra.*

[127] More generally, $P(X|E_1 \& E_2 \& \dots \& E_n \& E_{n+1}) =$
$P(E_{n+1}|X \& E_1 \& \dots \& E_n)/P(E_{n+1}|E_1 \& \dots \& E_n) \bullet P(X|E_1 \& \dots \& E_n)$.

[128] *See* note 75 *supra.*

[129] The crucial factor is not so much that of time sequence as that of priority in thought. Even if all of the evidence has in fact been introduced before the trier is asked to quantify the probative force of a limited part of it without taking account of the rest, the problems discussed here still obtain.

[130] *See* p. 1366 *supra.*

[131] Such facts would typically include, for example, hearsay of a sort that, though inadmissible at trial, may support a valid indictment. *See* Costello v. United States, 350 U.S. 359, 362-63 (1956). *But cf.* United States v. Payton, 363 F.2d 996 (2d Cir.), *cert. denied,* 385 U.S. 993 (1966).

[132] Nor would it be a satisfactory solution to instruct the jury to assign an artificially low starting value to P(X) on the pretense that the accused was plucked randomly from the population, even if the jury could be trusted to follow such an instruction. Suppose that P(X) would be thought to equal 1/2 but for the suggested pretense of random selection, and suppose that the prosecution's evidence *E* is so compelling that, given this P(X), it turns out that $P(X|E) > .999$. This would be the case, for example, if $P(E|X) = 1/2$ and $P(E\ |\text{not-}X) = 1/2{,}000$. *See* formula (5) at p. 1352 *supra.* Given this same evidence *E,* if the trier were to pretend that the accused had been chosen randomly from the population of the United States and were thus to treat P(X) as equal to approximately 1/200,000,000, he would obtain $P(X|E) < 1/200{,}000$--a miniscule, and evidently understated, probability that the defendant did the killing. If, however, the trier could manage to treat P(X) as *completely indeterminate* until at least some evidence *E'* (some significant part of *E*) had been introduced, he might rationally assign to P(X|E') a value high enough to bring P(X|E & E') close to .999 or higher. *See* note 125 *supra.* What this suggests is that setting P(X) equal to an artificially small (and essentially meaningless) quantity at the outset of the trial may distort the final probability downward in a way that need not occur if *no judgment at all* is made about the starting value of P(X) apart from at least some significant evidence in the case. *Cf.* note 100 *supra.* But once the jury is invited to assess the probability of guilt in light of less than *all* the evidence, there can be no assurance that it will not make an initial assessment that depends on *none* of the evidence. And, as has been shown, such an assessment can cause serious distortions whether the existence of an indictment or other charge is treated as probative background information or as equivalent to random selection. It should be noted that, without using Bayes' Theorem to compute the figures employed in this footnote, the degree of understatement that might be caused by an artificially deflated starting value for P(X) would be difficult if not impossible to assess, a fact that again serves to illustrate the usefulness of mathematical techniques in illuminating the process of proof even when, and perhaps especially when, one is rejecting the formal application of such techniques in the process itself. *See also* note 115 *supra.*

[133] Tribe, *An Ounce of Detention: Preventive Justice in the World of John Mitchell,* 56 U. VA. L. REV. 371, 404 (1970) [hereinafter cited as Tribe].

[134] *Id.* 404-06.

[135] *See* pp. 1372-75 & 1390 *infra.*

[136] The argument I advance here--unlike the arguments beginning at pp. 1358-59 *supra,* and unlike at least part of the argument beginning at p. 1368 *supra,* is meant to apply only to criminal cases, and perhaps only to those criminal cases in which either the penalty attached or the moral blame imputed makes the crime a sufficiently “serious” one.

[137] Although this notion is readily confused with the presumption of innocence, *discussed at* pp. 1370-71 *supra,* it is in fact quite different and rests on a partially overlapping but partially distinct set of objectives. To some extent, in fact, the concept that conviction is proper only after all real doubt has been dispelled may tend to undercut the purposes served by the presumption of innocence, for that concept suggests that a defendant's acquittal signifies only the existence of some doubt as to his guilt,

whereas one function of the presumption of innocence is to encourage the community to treat a defendant's acquittal as banishing all lingering suspicion that he might have been guilty. *See* p. 1370 *supra.*

[138] Even when the number measures only one element of the offense and omits an element like intent, *see* pp. 1365-66 *supra,* it sets an upper bound on the probability of guilt, and the argument made below follows a fortiori.

[139] Tribe 385-87.

[140] Even if this were the case, I would find it difficult wholly to ignore the fact that at least some who witness the trial process might interpret a series of publicly proclaimed decisions to condemn in the face of numerically measurable doubt as teaching that the sacrifice of innocent men is not to be regarded as a terribly serious matter. *See id.* 387 n.65. Those who adopt this interpretation might become more willing than they should be to tolerate the sacrifice of others and less confident than they ought to be of their own security from unjust conviction. Although such an interpretation would be in error, it would not be *wholly* unjustified, for a society that does not recoil from confronting defendants in quantitative terms with the magnitude of its willingness to risk their erroneous conviction *is,* it seems to me, a society that takes the tragic necessity of such sacrifice less seriously than it might.

When the Supreme Court held that "the Due Process Clause protects the accused against conviction except upon proof beyond a reasonable doubt of every fact necessary to constitute the crime with which he is charged," it stressed the importance of not leaving the community "in doubt whether innocent men are being condemned," reasoning in part that such doubt would dilute "the moral force of the criminal law" and in part that it would impair the confidence of "every individual going about his ordinary affairs ... that his government cannot adjudge him guilty of a criminal offense without convincing a proper factfinder of his guilt with utmost certainty." *In re* Winship, 397 U.S. 358, 364 (1970). If due process required less than this publicly announced insistence upon "a subjective state of certitude," *id., quoting* Dorsen & Rezneck, *In re Gault and the Future of Juvenile Law,* 1 FAMILY L.Q. 1, 26 (1967), the Court seemed to be saying, the sense of security conferred by a system that at least proclaims an unwillingness to punish in the face of palpable doubt would be irreparably eroded. *See* Tribe 389.

Both callousness and insecurity, then, might be increased by the explicit quantification of jury doubts in criminal trials--whether or not it would be *factually accurate* to describe the trial system as imposing criminal sanctions in the face of quantitatively measured uncertainty in particular cases. In considering a somewhat analogous problem in the area of accident law, Professor Calabresi has argued that there is a great difference in social cost between a set of individual market choices that indirectly sacrifice human lives by investing less than possible in life-saving resources and a collective societal choice that consciously and calculatingly sacrifices precisely the same lives for exactly the same reasons of economy. *See* Calabresi, *Reflections on Medical Experimentation in Humans,* 1969 DAEDALUS 387, 388-92. *See also* Schelling, *The Life You Save May Be Your Own,* in PROBLEMS IN PUBLIC EXPENDITURE ANALYSIS 127, 142-62 (S.B. Chase ed. 1968).

[141] *See The Supreme Court, 1969 Term,* 84 HARV. L. REV. 1, 157 & nn.8, 9 (1970).

[142] In some respects, it should be stressed, the insistence on certainty does *not* parallel the presumption of innocence. *See* note 137 *supra.*

[143] Tolerating a system in which perhaps one innocent man in a hundred is erroneously convicted despite each jury's attempt to make *as few mistakes as possible* is in this respect vastly different from instructing a jury to *aim* at a 1% rate (or even a .1% rate) of mistaken convictions. *See* Tribe 385-86, 388.

[144] Dershowitz, *Preventive Detention: Social Threat,* TRIAL, Dec.-Jan. 1969-70, at 22.

[145] *Contra,* Finkelstein & Fairley 504; Ashford & Risinger, *Presumptions, Assumptions and Due Process in Criminal Cases: A Theoretical Overview,* 79 YALE L.J. 165, 183 (1969); Broun & Kelly, *supra* note 5, at 27.

[146] This seems to me a much more plausible account of the fuzziness of the "reasonable doubt" concept than does the alternative account that "courts shun responsibility for fixing a more precise threshold probability because they feel it should vary to some extent from case to case." Cullison, *supra* note 5, at 567. *See also* Broun & Kelly, *supra* note 5, at 31; Kaplan 1073.

[147] Finkelstein & Fairley 517; *see* p. 1358 *supra.*

[148] Suppose, for example, that each of three items of evidence, E_1, E_2, and E_3, has the effect of increasing a prior 1 percent suspicion of guilt ($P(X) = .01$) tenfold, so that $P(X|E_1) = P(X|E_2) = P(X|E_3) = .1$ If E_1, E_2, and E_3 are conditionally independent of X and not-X, *see* note 75 *supra,* then it turns out that $P(X|E_1 \& E_2 \& E_3)$ is *in excess of .93,* a result that might be counter-intuitive for many laymen. For another illustration, *see* RAIFFA, *supra* note 58, at 20-21.

[149] Finkelstein & Fairley 517.

[150] The argument I am here advancing applies with greatest force in the criminal context, but it also has some significance in much ordinary civil litigation.

[151] Hart & McNaughton, *Evidence and Inference in the Law,* in EVIDENCE AND INFERENCE 48, 52 (D. Lerner ed. 1958). I do not exclude the possibility that, in extraordinary cases, and especially in cases involving highly technical controversies, the "historical" function may be so dominant and the need for public comprehension so peripheral that a different analysis would be in order, laying greater stress on trial accuracy and less on the elements of drama and ritual.

[152] *See* United States v. Spock, 416 F.2d 165, 182 (1st Cir. 1969).

[153] *See generally* A. ENGELMANN, A HISTORY OF CONTINENTAL CIVIL PROCEDURE 155, 651-52 (1928).

[154] *See* p. 1361 & notes 33 and 102 *supra.*

[155] Some, but by no means all, of the costs of precision identified in this section are primarily costs for persons actually or potentially accused of crime. To the extent that a defendant in a criminal case wishes to employ mathematical methods in his own defense, these costs obviously weigh less heavily than they do in the case of prosecutorial use. One can imagine a variety of defensive uses of mathematics--for example, to establish the likelihood of an "accidental" cause of a seemingly incriminating event (as in the parking case, p. 1340 *supra*), or to show the likelihood that a person other than the accused committed the crime (as in the police and mistress cases, p. 1341 *supra,* modified by assuming different defendants from those there posited). But the most common defensive use would probably be the translation into quantitative form of an expert's damaging opinion that a certain physical trace or combination of traces must "almost certainly" have been left by the accused. *See* Finkelstein & Fairley 517. Courts otherwise hostile to probabilistic proof have at times allowed such quantification of expert opinion about trace evidence even at the prosecutor's initiation. *See* People v. Jordan, 45 Cal. 2d 697, 707, 290 P.2d 484, 490 (1955); *cf.* Miller v. State, 240 Ark. 340, 343-44, 399 S.W.2d 268, 270 (1966) (quantification rejected only because the prosecutor laid "no foundation upon which to base his probabilities"). Although the analysis of the preceding section would make me somewhat reluctant to accept such holdings (particularly in light of the "selection effect" described in note 40 *supra*), I am of the tentative view that the criminal defendant should nonetheless be permitted to initiate the quantification of this sort of expert opinion in order to establish a "reasonable doubt" as to his guilt. And, once such quantification has been initiated by the defense, the case for allowing the prosecution to rebut in mathematical terms becomes quite persuasive.

[156] *See* Kaplan, *supra* note 5.

[157] *See* Cullison, *supra* note 5.

[158] *See generally* RAIFFA, *supra* note 58.

[159] It is not entirely clear to what extent the model is intended by Kaplan or Cullison as a description of how trial decisions are in fact made, *see, e.g.,* Kaplan 1069-70, 1075, to what extent it is offered as an heuristic device for illuminating the trial process, *see, e.g., id.* 1066, 1091, and to what extent it is meant as a normative guide to the trier's choice of verdict, *see, e.g., id.* 1065, 1072-74, 1092. I am concerned here with the model's heuristic and normative roles only, and of the four criticisms I later advance, *see* pp. 1381-85 *infra,* all but one, *see* p. 1384 *infra,* apply to the former as well as to the latter.

[160] It might be objected that the gambling analogy is a weak one insofar as the payoffs in the trial "game" accrue directly to persons other than the decision-maker. Since there are obviously significant psychological payoffs for the gambler as well, however, the objection seems to me a superficial one. *Cf.* note 60 *supra.*

[161] An alternative approach is possible, focusing simply upon the "disutility" of each of the two possible kinds of errors (erroneous conviction or erroneous acquittal), *see* note 168 *infra,* but expressing the problem in the terms used here facilitates the assignment of numbers to the various outcomes and makes somewhat more transparent the "cognitive dissonance" problem *discussed at* pp. 1383-84 *infra.*

[162] It should be noted that the procedural rule produced by the model will always have this numerical form; it can never assume the indefinite shape of "subjective certitude" or "guilt beyond a reasonable doubt." *Cf.* pp. 1374-75 *supra.* In his concurring opinion in *In re* Winship, 397 U.S. 358, 368 (1970), Mr. Justice Harlan argued that the "reasonable doubt" standard in criminal cases, like the quite different "preponderance of the evidence" standard in much civil litigation, merely reflects an assessment of the comparative social disutility of erroneous acquittal and erroneous conviction. *Id.* 370-71. Given the societal recognition that the latter error is far worse than the former, *id.* 372, a demanding burden of proof is imposed on the prosecution in order to assure that men are wrongly convicted much less often than they are wrongly acquitted. *Id.* at 371. *See also* Ball, *supra* note 37, at 816. This analysis, for which Mr. Justice Harlan credits Kaplan, *supra* note 5, 397 U.S. at 370 n.2, suffers from all of the defects I will shortly discuss with respect to Kaplan, *see* pp. 1381-85 *infra,* and suffers in addition from the defect that it proves too little. Specifically, the objective of assuring that erroneous acquittals of the guilty occur with greater frequency than erroneous convictions of the innocent demands only that the prosecution be required to prove its case more convincingly than must a civil plaintiff (*e.g.,* by "clear and convincing evidence," or perhaps "to a probability of 9/10"), *not* that it produce the "subjective state of certitude" stressed by the Court's majority opinion. *See The Supreme Court, 1969 Term,* 84 HARV. L. REV. 1, 158 n.13 (1970).

[163] *See* pp. 1372-74 *supra.*

[164] *See* note 140 *supra.*

[165] *See* Kaplan 1073.

[166] To illustrate the sharp difference between this view and the model put forth by Kaplan, note how Kaplan explains why our legal system typically excludes evidence of previous convictions from the prosecution's case-in-chief. This is done, he says, because including such evidence might lead the jurors "to the perhaps rational but clearly undesirable conclusion that because of his earlier convictions, D_i, the disutility of convicting the defendant should he be innocent, is minimal," Kaplan 1074, and that a low probability of present guilt should thus suffice to warrant his conviction. *Id.* 1077. But if it were simply a matter of fitting the standard of proof to the comparative utilities and disutilities of the four possible outcomes, *why* would that conclusion be "clearly undesirable"? Because, we are told, ours is "a system of justice that regards it as crucial that the defendant be found guilty only of the crime specifically charged." *Id.* 1074. Yet, if that is so, and if a conclusion that flies in its face nonetheless emerges as "perhaps rational" and indeed inevitable within the four corners of Kaplan's utilities, must one not conclude that the model built on those four utilities is inherently defective? Our system typically excludes prior convictions (with, of course, many exceptions) for the kinds of reasons that any adequate model of the criminal trial must somehow reflect--for reasons of repose; for the prevention of multiple punishment; for the appearance of fairness; for the preservation of a substantive system of law in which the accused, however long his record, can by his own choice avoid future entanglement with the criminal process, *cf.* Tribe 394-96; and for the preservation of a procedural system of law in which the accused, whatever his background, is given a well-defined opportunity to rebut a precise charge. *Cf. id.* 392-94. All of those factors enter into the question whether a man's trial was a "fair" one; none of them figures in the simplistic calculation of how desirable or undesirable would be each of its four possible outcomes.

[167] *See* note 140 *supra.*

[168] The Kaplan-Cullison model, to generalize the computation performed at pp. 1379-81 *supra,* yields the rule that conviction should be preferred to acquittal whenever *P,* the final probability estimate of the defendant's guilt, exceeds the quotient:

$1/1 + U(C_G) - U(A_G)/U(A_I) - U(C_I)$.

Kaplan designates the difference $U(C_G) - U(A_G)$ by the symbol D_g and the difference $U(A_I) - U(C_I)$ by the symbol D_i. *See* note 161 *supra.* If, as I have suggested in text, the values of $U(C_G)$, $U(A_G)$, $U(A_I)$, and $U(C_I)$ themselves depend on *P,* then the rule will have a more complex form. Thus, if the utilities or at least their differences ($U(C_G) - U(A_G)$ and $U(A_I) - U(C_I)$)

depend in a linear way on *P,* there will exist numbers U_1, U_2, and U_3 such that conviction is preferable to acquittal whenever $U_1P^2 + U_2P + U_3 > 0$. Because such an equation can have two roots between 0 and 1, there may be no single threshold value *P** such that conviction is preferable to acquittal for all $P \geq P^*$.

[169] A typical example would be the question whether or not to drill for oil at a given site before one's option expires, given incomplete information about such variables as the cost of drilling and the extent of oil deposits at the site. *See* RAIFFA, *supra* note 58, at *xx.*

[170] An analogous problem outside the trial context would be presented by a choice among alternative medical strategies, all entailing some risks, for a patient who might or might not have cancer. The application of classical techniques of decision-analysis to such situations, in which the decisionmaker cannot be entirely neutral with respect to the uncertain facts underlying his problem, is a matter of much current interest and research within the decision-analysis profession, although I am aware of no published discussions of the problem thus far.

[171] *See generally* L. FESTINGER, A THEORY OF COGNITIVE DISSONANCE (1957).

[172] *But see* note 177 *infra.* I would not have as much quarrel with techniques that called upon the factfinder to think in less formal terms about how high a probability of guilt to require as a precondition of returning a verdict against the defendant, but any method yielding a numerical conclusion to that question would be subject to the basic objection I have made to attempts at quantifying the final probability of guilt. *See* pp. 1372-75 *supra.*

[173] Kaplan 1075.

[174] The answer to such an obviously crucial question as “how much deterrent effect will flow from convicting whenever the probability of guilt exceeds 4/7” cannot influence the trier's assessment of the desirability or undesirability of correctly or erroneously convicting or acquitting any given defendant, and hence cannot affect the decision in the Kaplan-Cullison model whether to treat a 4/7 probability of guilt as sufficient to convict.

[175] I am much indebted to Professor Thomas C. Schelling for helping me translate my earlier and more intuitive formulation of this general approach into the one employed in the present article. Techniques of a related sort are employed by both Becker, *supra* note 5, and Birmingham, *supra* note 5.

[176] Of course it might turn out that the frequency of offenses depends strongly not only on the probability of conviction if guilty but also on the ratio of convictions to trials, or even on the absolute number of convictions. If this should prove to be the case, a rulemaker using this sort of approach might be tempted to design the system so as to convict more innocents, not only as an unavoidable cost of convicting a higher percentage of guilty, but also as part of a deliberate strategy of deterrence. Needless to say, this may well be a decisive objection to this form of analysis.

[177] If, as one might well suspect, the constituents who matter most to the rulemaker, *see* V.O. KEY, AMERICAN STATE POLITICS: AN INTRODUCTION 140-41 (1956), will themselves be insulated by social or economic status from the “insecurity costs” of a rising risk of convicting innocents, the rules arrived at through the procedure outlined here (and perhaps, though by no means certainly, the rules that would be arrived at in less calculating ways as well) will expose categories of persons more susceptible to false arrest and mistaken conviction to a greater danger of such misfortunes than would result from rules chosen by, or on behalf of, men acting under a veil of ignorance as to their ultimate status in the society they are designing. *See* Rawls, *Justice as Fairness,* 67 PHILOS. REV. 145 (1958). The jury, on the other hand, is--or can more readily be made into--a body of persons more likely than the typical legislator's important constituents to pay in false convictions the price of relaxed prosecutorial burdens. This may create a powerful reason for leaving to the jury broader discretion with respect to such matters as standards of proof than was argued for at p. 1384 *supra,* at least so long as the delegation of such discretion takes a sufficiently inarticulate form.

[178] Each rulemaker has, of course, an infinite number of these preference contours for a given crime; in the above illustration, only a representative subset of the entire family of contours can be depicted.

[179] To see why *Q* is optimal, simply imagine any alternative points on the choice set *C* such as *R* or *T.* Because *S, Q,* and *U* in the above illustration lie on the same preference contour or indifference curve, the rulemaker feels equally satisfied at those three points. But he clearly feels better at *S* than at *R* and better at *U* than at *T,* since in each case the first point of the pair provides the same benefit as the second (the same percentage of guilty convicted) at lower cost than the second (a lower percentage of innocent convicted), assuming that the perversity described in note 176 *supra,* does not obtain. Since the rulemaker is indifferent as among *S, Q,* and *U,* and prefers *S* to *R* and *U* to *T,* he must prefer *Q* to either *R* or *T,* which implies that *Q* is indeed optimal. It should be noted, however, that the ability of this model to yield a unique optimum depends upon several assumptions with respect to the shape of the preference contours, that of the choice set, and the relationships among them.

[180] *See* pp. 1358-77 *supra.*

[181] *Cf.* pp. 1361-66 *supra.*

[182] As, for example, by relaxing the privilege against self-incrimination, or the requirement of proof beyond a reasonable doubt.

[183] Particularly is this so in light of the expressive role of procedure discussed at pp. 1391-93 *infra.*

[184] *See* pp. 1370-75 *supra.*

[185] Although I regard this as an important problem, I do not think it quite as significant as the analogous problem in the context of mathematical proof, where the decision to take a visible and calculated risk of erroneously convicting a specific accused person is more dramatic, may be thought to entail a lack of respect for the accused as an individual person, and seems more likely to have wide-ranging psychological impact. *See* pp. 1372-75 *supra. Cf.* Tribe, *supra* note 133, at 386 n.65.

[186] *See, e.g.,* Dworkin, *The Model of Rules,* 35 U. CHI. L. REV. 14 (1967). *See also* p. 1375 *supra* and note 162 *supra.* This is not to deny, of course, that "bright line" rules are occasionally preferable, *see, e.g.,* Bok, *Section 7 of the Clayton Act and the Merging of Law and Economics,* 74 HARV. L. REV. 226, 270-73, 350-55 (1960), but only to stress the importance of being able to choose general principles when they seem better suited to one's purposes.

[187] *See generally* C. LINDBLOM, THE INTELLIGENCE OF DEMOCRACY 207-08 (1965) (a study of decisionmaking through mutual adjustment).

[188] *See, e.g.,* Freund, *Privacy: One Right or Many,* 13 NOMOS 182 (1971).

[189] *See* pp. 1370-71 & 1372-76 *supra.*

[190] *See, e.g.,* C. FRIED, AN ANATOMY OF VALUES 125-32 (1970); E. GOFFMAN, INTERACTION RITUAL 10-11, 19, 54 (1967). *See also* J. Feinberg, *The Expressive Function of Punishment,* 49 THE MONIST 397 (1965), in DOING AND DESERVING (1970). For much of the discussion that follows, I am heavily indebted to the work of both Professor Goffman and Professor Fried.

[191] Staub v. City of Baxley, 355 U.S. 313, 320 (1958).

[192] As Thurman Arnold once observed,

> Trials are like the miracle or morality plays of ancient times. They dramatically present the conflicting moral values of a community in a way that could not be done by logical formalization. Civil trials perform this function as well as do criminal trials, but the more important emotional impact upon a society occurs in a criminal trial.

Arnold, *The Criminal Trial as a Symbol of Public Morality,* in CRIMINAL JUSTICE IN OUR TIME 141-42, 143-44 (A. Howard ed. 1965).

[193] *See* pp. 1370-71 *supra.*

[194] *See, e.g.,* Gideon v. Wainwright, 372 U.S. 335 (1963).

[195] *See, e.g.,* Pointer v. Texas, 380 U.S. 400 (1965).

[196] *See, e.g.,* Malloy v. Hogan, 378 U.S. 1 (1964).

[197] *E.g.,* the defendant's right in some circumstances to exclude evidence of prior crimes, discussed in note 166 *supra.*

[198] *E.g.,* fewer erroneous convictions.

[199] *See, e.g.,* pp. 1370-71 & 1373-75 *supra.*

[200] *See, e.g.,* Nozick, *Moral Complications and Moral Structures,* 13 NAT. L.F. 1, 3 (1968), *discussed in* C. FRIED, AN ANATOMY OF VALUES 95, 157 (1970).

[201] For example, the illustration given by Nozick, *id.,* proposes that for any theory *T* that describes which actions are morally impermissible, one may define a function *f* whose maximization mirrors the structure of *T* by setting f(A) = 0 whenever action *A* is impermissible according to *T* and f(A) = 1 otherwise. But if some *A*'s have the perverse effect of *changing T,* then even this "gimmicked-up" real-valued function will not do as a function whose maximization mirrors the moral values implicit in *T.* The character of at least some procedural rules, I am suggesting, is related to our value system precisely as *T*-changing *A*'s are related to *T.*

[202] Even if one takes the view that means and ends (or values) differ not in kind but only in degree, this argument still has significance as indicative of how extraordinarily little can be taken as "given," and hence as subject to weighted maximization, in the procedural area.

[203] *See* p. 1377 *supra.*

84 HVLR 1329

End of Document

UNITED STATES DISTRICT COURT
WESTERN DISTRICT OF MICHIGAN
SOUTHERN DIVISION

UNITED STATES OF AMERICA,

Plaintiff,

v.

DANIEL GISSANTANER,

Defendant.
/

Case no. 1:17-cr-130

Hon. Janet T. Neff
United States District Judge

Hon. Ray Kent
United States Magistrate Judge

DEFENDANT'S SUPPLEMENTAL BRIEFING REGARDING MOTION TO EXCLUDE DNA EVIDENCE

NOW COMES, the defendant, Daniel Gissantaner, by and through his attorney, Joanna C. Kloet, Assistant Federal Public Defender, and hereby files this supplemental brief, and Attachments 1, 2, and 3 as the three documentary exhibits it considers the strongest support for his position, in support of his Motion to Exclude DNA Evidence.

I. STRmix is unreliable because it fails to adhere to established software standards and has not been adequately validated for complex mixtures.

The Government has the burden of proving that contested, novel scientific evidence is both relevant and reliable. *Daubert v. Merrill-Dow Pharmaceuticals, Inc.*, 509 U.S. 579, 589 (1993). A "key question to be answered in determining whether a theory or technique is scientific knowledge that will assist the trier of fact will be whether it can be (and has been) tested." *Id.* at 593. Although the subject of scientific testimony need not be known to a certainty, "an inference or assertion must be derived by the scientific method," and "[p]roposed testimony must be supported by appropriate ***validation*** – i.e., 'good grounds,' based on what is known." *Id.* at 590 (emphasis added). This requirement "establishes a standard of evidentiary reliability." *Id.* Here, STRmix fails minimum standards for software validation, and furthermore has not been sufficiently validated for complex mixtures of three or more contributors with a minor contributor of only 7%.

A. STRmix's validation failed to adhere to controlling software standards.

For decades, the field of software engineering has "described appropriate approaches to software development processes and codified them in centralized standards documents." (Attachment 1, p 33.) These standards provide a "baseline to adhere to" and are recognized at national, international levels." (ECF No. 78, PageID.2868-69.) These principles ensure efficiency, productivity, and especially, protect against defects that have significant consequences. (ECF No. 78, PageID.2881.) The lack of implementation of software engineering standards have led to significant catastrophes, including the death of several patients receiving Therac-25 radiation therapy as well as the failure of NASA's Mars Climate Orbiter. (Attachment 1, p 10.) Insufficient testing processes can affect the quality of the software, which in turn can affect the final calculation of the LR in a way that is prejudicial to an accused in a criminal case. (ECF No. 78, PageID.2822.)

Federal and state government agencies, including the FDA, the Department of Defense, the Nuclear Regulatory Commission, the Michigan Department of Technology, Management, and Budget, and the FBI's Combined DNA Indexing System, cite the Institute of Electrical and Electronic Engineers (IEEE) standards as authoritative documents. (Attachment 1, pp 8-9.) IEEE "is the most commonly used standard setting body in computer science." (ECF No. 77, PageID.2564.) These standards apply broadly, and can be used to validate smartphone apps, navigation systems in airplanes, video games – in fact, "anything that has a software component, that component can be managed with an IEEE standard process." (ECF No. 78, PageID.2870.)

Important or high-risk software requires more robust validation. A "core feature" of software validation under IEEE standards is the ***integrity level***, which "quantifies the complexity, criticality, risk, safety level, security level, desired performance, reliability or other system-unique characteristics[.]" (Attachment 1, p 13.) For example, an airline navigation system "is likely to have a different integrity level than a game on your Smart phone." (ECF No. 78, PageID.2876.) This determination is based on the consequences of a software malfunction as well as the likelihood of failure. (Attachment 1, p 14.) ***Severity*** is ranked in terms of the degree of the effects of failure, which is "often described in terms of financial loss or damages, loss of human life or physical, physical damage to a person or property," in reference to the stakeholders, who are those affected by the operation of these systems. (ECF No. 78, PageID.2874, 2876-77.) These considerations can be used to decide if a program should be adopted, how it should be adopted, and whether "more thorough evaluations" are needed. (ECF No. 78, PageID.2874-75.)

Rigorous validation requires "the identification and execution of a number of tasks intended to inspect and evaluate that system." (ECF No. 78, PageID.2867-68.) Software validation entails making sure the software behaves as intended—to ensure "that nothing got lost or improperly introduced during the translation from concept to program." (ECF No. 78, PageID.2880.) Computer scientists expect rigorous software development practices to require and generate "a long list of materials" during the process, reflecting risk analysis, requirements and specifications, and issue tracking. (ECF No. 78, PageID.2879.) Requirements are "a list of behaviors that [the program] is supposed to adhere to," and specifications are a "translation of requirements to more technical notation," which represent the detailed "behaviors of the system before any code actually gets written." (ECF No. 78, PageID.2879.)

However, few of these principles were employed in STRmix version 2.3.07. (ECF No. 78, PageID.2897.) First, no integrity level assessment or risk analysis was undertaken. (ECF No. 78, PageID.2898; Attachment 1, p 33.) Next, "[n]o description of requirements[,] specifications, or a clear mechanism to derive such specifications, appears to exist," such that during his review of the program, defense expert Nathan Adams was unable to evaluate "test coverage and scope, what's appropriate to test, which components were tested, how they were tested, the overall order of the code[.]" (ECF No. 78, PageID.2808-99, 2908; Attachment 1, p 33.) Adams testified that in the context of software utilized for forensic DNA mixture interpretation, important characteristics of forensic DNA mixture interpretation, such as stutter artifacts, should be codified in the requirements document and tested against when the software has been constructed." (ECF No. 78, PageID.2892.)

Additionally, the software's source code indicates the program lacks automated software testing. (Attachment 1, p 33.) In fact, "no test package exists for the DyNAmix package, which contains much of the core functionality of the STRmix system." (Attachment 1, p 23.) The "apparent lack of automated software testing for the DyNAmix package suggests that all testing is based on manual inspection," and furthermore, is limited. (Attachment 1, p 33.) Importantly, although Dr. Buckleton asserted that "at least two groups have done some quite extensive by hand verification of the STRmix outputs," he conceded that "you can't even begin to do all" of the calculations. (ECF No. 77, PageID.2590-91.) An automated software testing package would have addressed these immense gaps in verification. However, STRmix lacks any "indication that an attempt has been made to generate an objective, quantified description of test coverage." (Attachment 1, p 33.)

Contrary to Dr. Buckleton's claims, some errors disclosed during his STRmix training were not listed publicly. (ECF No. 78, PageID.2904.) During his review, Adams also uncovered a document listing code changes that suggested changes were made to latent defects that were not made publicly available. (Attachment 1, pp 27-28.) Finally, Adams observed at least one other unpublished problem, which warned of the possibility that STRmix may not consider all possible genotypes at all loci. (ECF No. 78, PageID.2905-06; Attachment 1, pp 28-29.) This issue could affect weights of the genotypes that STRmix did consider, which in turn, will affect the LR generated. (ECF No. 78, PageID.2906.)

The existence of errors not published by STRmix is concerning, because "[t]he dissemination of inappropriate or knowably incorrect behavior is something that is important to not just the users of the software but anybody who has an interest in the software," and "[p]roblems anywhere throughout the software development process can affect the final conclusion reported by the software." (ECF No. 78, PageID.2913-14.) MSP's Jeffrey Nye testified that when he purchased the software, "there wasn't as much information out there about like coding errors and things of that nature." (ECF No. 77, PageID.2692.) Furthermore, although Dr. Buckleton claimed that only one "miscode" affected the calculation of the LR in version 2.3.07 of STRmix, Adams testified that version 2.3.07 was affected by most of these defects. (ECF No. 77, PageID.2574; ECF No. 78, PageID.2901-02.)

In his review of STRmix, Adams also did not see any reference to issue or "bug" trackers, which are "common frameworks for monitoring the identification and mitigation of defects" or changes. (Attachment 1, p 20.) Although he saw passing reference to a Github repository, he explained that Github is "primarily known as a version control system," and issue trackers and version control systems, while related, are "definitely discrete topics." (ECF No. 78, PageID.2663; Attachment 1, p 20.) For his part, Dr. Buckleton appeared to be unsure about the very purpose of Github, testifying that he "think[s]" issue trackers "relate to the use of Github." (ECF No. 77, PageID.2567.) In any case, Dr. Buckleton conceded that he did not know whether he could "technically" be given access to the information that reflects the "entire back history of STRmix." (ECF No. 77, PageID.2567.)

Next, STRmix showed a lack of orderliness and consistent authorship throughout the code, "suggestive of a difficult to maintain code base." (ECF No. 78, PageID.2908-09.) Portions of the source code appeared to have been written by the STRmix team, while others appeared to have been written by outside parties, while still others "simply have no indication as to who wrote them, when they were written, when they were modified, how they were modified." (ECF No. 78, PageID.2908-12; Attachment 1, p 30.) The code created by the in-house STRmix team contained cluttered, commented-out, nonfunctional code, a feature not expected in a well-maintained, professionally developed software product. (ECF No. 78, PageID.2908-10; Attachment 1, p 31.) Notably, the segments of the code that appeared to be developed by outside parties did not have these issues. (ECF No. 78, PageID.2910.)

"[T]he author of nearly all the code or possibly even all the code is Dr. Duncan Taylor." (ECF No. 77, PageID.2571.) Dr. Taylor is not a professional coder or computer scientist, but a forensic biologist "who programs." (ECF No. 77, PageID.2573.) Ultimately, Dr. Buckleton conceded that Adams' concerns about code quality "has some veracity," and assured the Court that the newest version of STRmix, written by professional coders, "would meet many of Mr. Adams's requirements now." (ECF No. 77, PageID.2564, 2574.) Software coding errors "have to be found by testing," and in fact, during the development of the upcoming version of STRmix, version 2.6, coded by professionals, 416 bugs have been identified. (ECF No. 77, PageID.2574-75, 2586-87.) Notably, Dr. Buckleton did not know how many miscodes were identified during testing of STRmix version 2.3.07. (ECF No. 77, PageID.2587.)

STRmix version 2.3.07 also lacks consistent use of object-oriented principles (OOP) in its overall design. (Attachment 1, pp 30-31.) For instance, code duplication, which "complicates testing and maintenance," is an activity opposed by OOP, but nonetheless was "frequent and widespread" in version 2.3.07. (Attachment 1, p 31.) Foundational OOPs, such as "inheritance, delegation, and reuse[,] appear to be avoided" within the DyNAmix section of STRmix. (Attachment 1, p 31.) Dr. Buckleton conceded that this particular criticism was "possibly with some justification" and declared that STRmix has since "upped the use" of OOPs in subsequent versions of STRmix. (ECF No. 77, PageID.2571.)

Adams opined that ***STRmix should not be relied upon*** "until additional methods of software quality assurance have been undertaken." (ECF No. 78, PageID.2939.) Adams concluded:

> Given these departures from basic software engineering practices, STRmix v2.3.07 should not be considered capable of being verified and validated against objective criteria. Accordingly, its calculations of likelihood ratios for complex mixtures (such as the profile generated during the course of testing item 2Ax in this case) should not be relied upon as accurate or 'operationally' correct. (Attachment 1, pp 33-34.)

The Government relies upon purported compliance with the FBI's SWGDAM guidelines for validating probabilistic genotyping systems as evidence that STRmix has been sufficiently validated. However, these SWGDAM guidelines "don't have an emphasis or even really a mention of software development principles." (ECF No. 78, Page ID.2883.) Thus, SWGDAM's recommended protocols for probabilistic genotyping systems does not contain principles that are specific to software, or that incorporate the principles of software validation that apply to software in other fields that are not related to forensic science. Unlike testing a new automobile by seeing if it arrives safely at a given destination, with a software that generates an LR, "no innate ground truth" exists, such that one could know that "when we arrive there[,] our process has appropriately worked." (ECF No. 78, PageID.2867.) Validation processes for software programs are qualitatively more important for systems that have "difficult to assess outputs" like a LR. (ECF No. 78, PageID.2867-68.) Notably, the International Society of Forensic Geneticists (ISFG) and the United Kingdom Forensic Science Regulator (UK FSR) have both recommended the incorporation of IEEE standards for testing probabilistic genotyping systems. (ECF No. 77, PageID.2565, 2589; ECF No. 78, PageID.2890.)

In sum, STRmix fails many important IEEE standards, standards that apply to all software. The *Daubert* standard requires that a proffered technology must be supported by ***appropriate*** validation. The importance of adhering to principles and standards for validation of software cannot be overstated where application of the software does not have a known ground truth. A lower standard for testing should not apply to forensic software that is used to convict, and therefore potentially incarcerate or exterminate, human beings. The evidence suggests the creators of STRmix 2.3.07 either were unaware of or openly rejected the application of well-established principles of software testing. Even if their reasons are attributable to something as benign as unfamiliarity with these well-established principles, abandoning them is not justifiable and is inconsistent with *Daubert*'s standards.

B. STRmix has not undergone adequate independent testing on complex mixtures of three or more where a minor donor comprises only 7%.

As PCAST observed, empirical testing of probabilistic genotyping systems has largely been limited to a narrow range of parameters, and further testing has been recommended. PCAST set forth limitations on the validity of probabilistic genotyping systems up to three-person mixtures in which the minor contributor is at least 20%. (ECF No. 77, PageID.2576-77.) Here, MSP forensic analyst Amber Smith testified that if the defendant was present in the mixture, he would be associated with the 7% contributor. (ECF No. 77, PageID.2724.) Thus, the mixture here is precisely the kind for which PCAST recognizes this software has not yet been adequately empirically tested. The LR in this case was generated by a methodology that extends beyond the accepted edges of reliability.

The Government relies on STRmix's developmental validation process and the local laboratories' internal validation studies to argue that STRmix has been adequately validated. (ECF No. 77, PageID.2578.) However, principles of sound science and common sense dictate that validation is not a task that can or should be entrusted to either the product's creator or its end-user. For instance, it would not be prudent to rely on the automaker's self-interested assurances that the product has passed the standards created by and executed by the automaker itself. Nor should an automobile manufacturer rely on its consumer-drivers to conduct the safety testing on a new car. Ensuring a car's safety, and its adherence to safety standards, is most appropriately a non-delegable responsibility of an impartial third party, such as the National Highway Traffic Safety Administration.

Under IEEE standards, ***independence*** is key component of software validation. (Attachment 1, p 15; ECF No. 78, PageID.2877-78.) Three distinct dimensions of independence are set forth by IEEE: (1) managerial independence, which ensures that there is not a pressure "exerted on whoever is reviewing the product," or pressure "to accept this as the way it is because your boss wants you to;" (2) financial independence, to avoid "the potential for a conflict of interest in that [a reviewer] might be hesitant to call out particular deficiencies of the software development process if their paycheck could be affected by it;" and (3) the technical component of independence – that is, independence between the reviewer and those who developed a component of the program, to avoid "biases or favoritism to particular components of the development process." (ECF No. 78, PageID.2877-78; Attachment 1, pp 15-16.)

The Government's Exhibits 4 and 5, which present the internal validations of 31 state and local laboratories and an internal validation study performed at the FBI, respectively, lack the requisite independence. (ECF No. 77, PageID.2547-49, 2579.) No testimony indicates that employees or affiliates of STRmix were not involved in the validation. In fact, the company that creates STRmix maintains a staff of at least 12 individuals that assist local laboratories who purchased STRmix. (ECF No. 77, Page ID.2555.) Dr. Buckleton testified that sometimes he is directly involved in these validations, and that he personally conducted two STRmix trainings for the Michigan State Police (MSP). (ECF No. 77, PageID.2554-55.) Tellingly, both Exhibits 4 and 5 list Drs. Buckleton, Taylor, and Bright, the STRmix creators, as authors. Likewise, closer inspection of the articles listed at Exhibit 12, which Dr. Buckleton described as a list of 33 peer-reviewed publications that he stated were "pertaining to STRmix" (in an unspecified way), reveals that each publication was authored in part by him, Dr. Bright, and/or Dr. Taylor. (ECF No. 77, PageID.2549.) Finally, Exhibit 23 is another article authored by Dr. Buckleton, which sets forth recommendations for validation of probabilistic genotyping programs. (ECF No. 77, PageID.2549.) Not surprisingly, Dr. Buckleton declared that STRmix met the requirements for validation set forth in this article. (ECF No. 77, PageID.2549.) Dr. Buckleton's testimony indicates that almost all of the peer-reviewed articles are published in one journal, Forensic Science International: Genetics, the journal he "tend[s] to submit to." (ECF No. 77, PageID.2587-88.)

When asked about independent verification by the Court, Dr. Buckleton agreed that it "should be separate from developers altogether."(ECF No. 77, PageID.2590.) However, when it came to STRmix, he could only muster that "to some extent we have met that as well," offering merely that "the group who do validation and verification do not do the programming." (ECF No. 77, PageID.2590.) And he immediately conceded that "[t]hat independence has softened" because they "find bugs" and the parties then communicate with one another. (ECF No. 77, PageID.2590.) Further, Dr. Buckleton's candid admission that he is "very proud" of the version of STRmix at issue in this case, reveals his personal bias towards a system that he himself produced and tested. (ECF No. 77, PageID.2526.) As an employee of ESR, and especially as the self-described "father of probabilistic genotyping," the involvement of Dr. Buckleton or other STRmix-affiliated individuals in these studies, stands in stark contrast to the principles of sound, independent validation, such as those set forth by IEEE. (ECF No. 77, PageID.2522-23.) Additionally, the internal validation studies bear another imprimatur of partiality – specifically, law enforcement agencies completing internal validation are not evaluated for their allegiance to science. Rather, they are evaluated on their ability to close cases and win convictions. This is an especially important consideration where, as here, the testing, steered by STRmix staff, did not involve a duplication of the tests performed on STRmix. (ECF No. 77, PageID.2620-2621.) To say these validation studies constitute adequate empirical testing defies both good science and good sense.

Furthermore, the tests performed during those validations are superficial and therefore inadequate. Evidence may be admitted only, if in addition to the reliability of the theory and process in general, the process is reliable when applied to the specific issue about which the expert is proposing to testify. *See Kumho Tire Co v. Carmichael*, 526 U.S. 137, 153 (1999). General acceptance of the reliability of a methodology does not compel a finding of reliability for any application of that methodology. Here, the application of STRmix to a complex mixture of at least three contributors of touch DNA, with a minor 7% contributor, is not reliable.

Although the authors of the study submitted as the Government's Exhibit 5 declare that "STRmix is sufficiently robust for implementation in forensic laboratories," purportedly based on running tests on "more than 300 samples," Table 1 of that article reveals that ***only about half*** of those mixtures contained at least three contributors. In addition, this study was based on simulated forensic specimens, not the damaged, degraded DNA mixtures encountered in the real world. As MSP's Nye testified, using lab-created samples in validation is "not always the best reflection of a sample type that we would see from a crime scene," and dirty or degraded evidence from a crime scene "represent[s] a better range[.]" (ECF No. 77, PageID.2672.) Yet Nye also identified a parallel concern with respect to using adjudicated samples for validation testing, admitting that "if you don't create the sample yourself, you really don't know what the ground truth of the sample is." (ECF No. 77, PageID.2672.) In other words, because the correct answers are not known, an error rate cannot be inferred. Whether a program has a known or potential rate of error is an important consideration for reliability under *Daubert*. *Daubert*, 509 U.S. at 594. Adherence to testing principles becomes even more critical for systems that have "difficult to assess outputs" like a LR. (ECF No. 78, PageID.2867-68.)

Finally, ensuring robust validation is even more difficult when, as here, full access to the program is not readily accessible. STRmix-sponsored training and the validation studies themselves did not involve a review of the source code. (ECF No. 77, PageID.2591, 2621.) Thus, lab directors, analysts, and other scientists who reviewed STRmix could not scrutinize and identify areas in the program where an issue might exist that could affect the reliability or accuracy of the LR. But as Nye testified, "it's not enough to just have a piece of software where you can click a button and get an answer out the other side." (ECF No. 77, PageID.2665.)

II. The LR generated by STRmix is not relevant because it is highly variable, such that the LR figure is uninformative yet unduly prejudicial.

An expert may testify to scientific knowledge provided it "will help the trier of fact." FRE 702. However, even relevant evidence can be prohibited under FRE 403 if its prejudicial effect outweighs its probative value. "Expert evidence can be both powerful and quite misleading because of the difficulty in evaluating it." *Daubert*, 509 U.S. at 595 (citations omitted). Accordingly, in FRE 403 analyses, a judge "exercises more control over experts than over lay witnesses." *Id*. Here, the LR lacks the necessary context to be useful to a jury. A LR represents only one of many possible reasonable interpretations of the data, yet it is presented to the jury as an objective certainty. Without knowing where the LR falls in the range of possible interpretations, the LR cannot help the jury and is overwhelmingly prejudicial.

Given the same input data, including the same parameters, the same number of contributors, and using the same population genetics data, STRmix will not generate the same answer a second time. (ECF No. 77, PageID.2550.) The Government may argue that this is necessarily fatal because the possible range of outcomes ostensibly are limited to one order of magnitude. In other words, a result indicating a LR of 49 million to one actually represents a range between 4.9 million to one and 490 million to one. However, the incurable problem is that even this broad range represents only a ***fraction*** of the possible LRs that can be obtained from the DNA mixture. And unlike the variance produced by STRmix's use of random numbers, no way of measuring this variance exists. In fact, the evidence suggests that it is likely at least several orders of magnitude.

As demonstrated at the hearing, MSP made a litany of subjective choices in determining the LR, including removal of spikes, pull up, and forward stutter. (ECF No. 77, PageID.2537, 2638.) The latter of these, "stutter," is a possible byproduct of the amplification process that occurs before the electropherogram is generated. (ECF No. 77, PageID.2610.) Although Dr. Buckleton claimed that STRmix has "mechanisms in place to avoid" this problem, he also admitted that the version of STRmix used in this case ***did not account for the possibility of forward stutter.*** (ECF

No. 77, PageID.2611, 2630.) Meaningfully, he stated that that deficiency has been fixed in later versions. (ECF No. 77, PageID.2611.)

Another example of subjectivity is when Ms. Smith threw out the data at locus D8 for what she deemed was oversaturation, even though the allele was ***under*** the saturation threshold of 25,000 RFU. (ECF No. 77, PageID.2710; ECF No. 78, PageID.2743-44.) She testified that her decision was a matter of "personal preference," and conceded that oversaturation was merely a possibility, not a certainty. (ECF No. 77, PageID.2710-14.) Importantly, this discarded information suggested four contributors to the mixture. (ECF No. 78, PageID.2744.) Although Ms. Smith explained that although she "feels" it is a three-contributor sample, she admitted that four contributors is "potentially correct," and that "there could be I guess four contributors there based on the artifacts that are present." (ECF No. 78, PageID.2744; ECF No. 77, PageID.2715.)

These types of calls are not "standard decisions that are being made in forensic DNA typing since its inception in the mid-90s"—instead, a STRmix analysis uses a new set of rules that are not applied in a standard DNA analysis. (ECF No. 77, PageID.2537, 2709.) The first MSP analyst, who was not trained on STRmix, had determined that "there was nothing wrong with the data that was detected in D8." (ECF No. 77, Page ID.2709.) But Smith explained that, as an analyst trained on STRmix, she is permitted to "look at certain areas more in-depth," and concluded the oversaturation occurred. (ECF No. 77, PageID.2709-10.)

Dr. Buckleton claimed that changing the number of contributors "does not materially affect the result and if it does, it affects it in the conservative direction," assuring the Court that "there is nothing bad that can happen" if the number of contributors inputted is incorrect. (ECF No. 77, PageID.2638.) However, his testimony is contradicted by Amber Smith's testimony that changing the number of contributors could potentially increase or reduce the LR. (ECF No. 78, PageID.2747.) Moreover, several research studies repudiate Dr. Buckleton's platitudes, demonstrating that varying the number of contributors in a probabilistic genotyping software program can affect an LR potentially to the detriment of the defendant. (ECF No. 78, PageID.2955.) For example, the authors of a recent study, attached hereto as Attachment 2, "examined the effect of incorrectly estimating the number of contributors for high order DNA mixtures with no or little drop-out." (Attachment 2, p 98.) They concluded that "large effects on the LR were obtained" if an incorrect number of contributors was assumed, and that "the LR varied considerably when the hypotheses used an incorrect number of contributors." (Attachment 2, pp 89, 92.) The authors explicitly observed that "***[a]ssuming an incorrect number of contributors may result in an inflated LR in favor of the prosecution.***" (Attachment 2, p 98 (emphasis added).)

Furthermore, the Gelman-Rubin diagnostic—a measure to determine whether STRmix has sufficiently analyzed the LR—in this case was over 1.2, the "limit many people would use." (ECF No. 77, PageID.2618-19.) Although Dr. Buckleton characterized 1.2 as a "soft threshold," he also testified that "often more complex mixtures do have higher" Gelman-Rubin numbers and that it "could have run for longer." (ECF No. 77, PageID.2632-33.) The choice to use the LR despite the high diagnostic is yet another "judgment call" made by the analyst. (ECF No. 77, PageID.2632.) Thus, the LR in this case is the product of many subjective decisions.

Additionally, the STRmix parameters vary greatly from laboratory to laboratory based on each individual lab's internal study. Ultimately, "nearly all" of these factors have an impact on the STRmix analysis. (ECF No. 77, PageID.2601.) For instance, the technology is so sensitive that the STRmix program attempts to account for some possible contamination in the laboratory. (ECF No. 77, PageID.2598.) Specifically, the "drop-in" parameter in STRmix tries to account for possible contamination in the lab by using "a model of alleles" that are "snowing from the ceiling" and may fall into a testing area. (ECF No. 77, PageID.2598-99.) Here, the drop-in cap at MSP was set at 400, meaning peaks below 400 would be potentially considered drop-in (contamination), and therefore not true allelic information from the original sample. (ECF No. 77, PageID.2601.) This means that using the same input on a different lab's STRmix would result in a different LR.

The extent to which the range of possible outcomes can vary, based on subjective analyst choices and individual lab parameters, is unknown, because STRmix has not attempted such a study. However, studying other probabilistic genotyping systems indicates that the potential variance is at least multiple orders of magnitude. Attachment 3 demonstrates the great variability among LRs generated by the same program. (Attachment 3; ECF No. 78; PageID.2770-71.) This study, published in Forensic Science International: Genetics, reported an "in[t]er lab trial" where different labs were asked to evaluate the same sample using probabilistic genotyping software. (ECF No. 78, PageID.2770.) The goal of this study was to determine "what's the variability or the range of interpretations from these different organizations." (ECF No. 78, Page ID.2770.) This study mostly addressed different laboratories' application of a probabilistic genotyping software program called LRmixStudio. (ECF No. 78, PageID.2770-71.)

This study revealed a huge range of results in the same program using the same underlying data: LR ranging from "2600 [to one] to something that's well beyond a billion, into the trillions [to one]." (ECF No. 78, PageID.2770-71; Attachment 3, p 161.) Put in terms of orders of magnitude, even looking at just results from LRmixStudio, the range of LR values covers up to 12 orders of magnitude. (Attachment 3, p 161.) The authors note that this vast range is "surprising" and states that the "[p]lausible explanations would range from the final allelic composition of the edited mixture profile [i.e., the results of the analyst's subjective decisions] to the parameters used in the software [i.e., the parameters determined by each individual lab]." (Attachment 3, p 161.)

Thus, the combination of the analyst's subjective decisions and the lab's individual parameters can have an immense effect on the LR.

Moreover, this wide variance only takes into account the variability within the same model. Models, or in this context, probabilistic genotyping software programs, attempt to represent the behavior of many different aspects of a system in different ways. Unlike some other models, STRmix discards certain information that could materially affect the LR—namely, if a peak is below the analytical threshold set by the laboratory, it is not used in the analysis and "will be thrown out." (ECF No. 77, PageID.2595-96, 2628-29.) However, low peaks could be evidence of the presence of true DNA, such that throwing out some of the data could lead potentially to an underestimation of the number of contributors. (ECF No. 77, PageID.2595-96.) In other words, several unknown and subjective variables, all of which can affect the output, are folded into a STRmix analysis that generates a single figure that that is presented to the jury without context.

For a result to be helpful to the jury, the jury must have some idea about the variability across different models as well. The *People v. Oral Hillary* case in the state of New York demonstrates that different models can lead to different results. In that case, TrueAllele—another widely used probabilistic genotyping system—rendered inconclusive results while STRmix produced a LR of 300,000 to one that the defendant's DNA was in the mixture. (ECF No. 77, PageID.2597-98.) Similarly, if a reasonable system incorporated PCAST's determination that for three-person mixtures the person of interest must make up at least 20% of the profile to produce a valid result, the system would necessarily reach an inconclusive result in the present case.

Even if we assume that STRmix is the best model for probabilistic genotyping, that does not mean that it is the only reasonable model. "Even career statisticians cannot objectively identify one model as authoritatively appropriate." (ECF No. 78, PageID.2801.) As Dr. Steven Lund, a statistician at the National Institute of Standards and Technology, testified, "whatever data you have there's a range of reasonable interpretations and you don't know what that range is until you've studied it from different perspectives." (ECF No. 78, PageID.2807-08.) Thus, in order to be helpful to a jury, a "systematic exploration of what the range of reasonable results might be for a given set of data" is a necessary prerequisite. (ECF No. 78, PageID.2764.)

For example, imagine a patient receives a cancer test. Regardless of the result, before making any decision about treatment, she would want to know the variability of the results. In other words, she would want to know where the results fell in the range of reasonable interpretations of the data. If the doctor's subjective choices could have a large effect on the results, the patient would want to know. Even if the patient generally trusts her doctor, knowing whether the doctor's choices determined the result is vital to evaluating the objective value of the test results. Likewise, if there are multiple tests that are used in the medical field, she would want to know how likely it is that a different reasonable test would give a different result. A monumental difference exists between receiving a result and knowing that 90 out of 100 other reasonable interpretations reach the same result, and knowing that 49 out of 100 reasonable interpretations reach the same result. If too much variability exists among the reasonable interpretations of the data, any single interpretation loses its objective value.

Applying this analysis to the present case, the LR produced by MSP is not helpful to the jury. Simply by virtue of STRmix's inherent randomness, the LR in this case may fall somewhere between 4.9 million to 490 million. The jury will not have any idea how much the analyst's subjective choices—and the individual lab parameters—further broaden this range. Similarly, the jury will have no idea where this result would fall among the outputs of other reasonable models: i.e., other probabilistic genotyping software programs. Accordingly, because the jury has no way to know whether this result is an outlier or in the heartland of reasonable results, it is not helpful.

Moreover, as the Court noted at the hearing, this LR would be incredibly prejudicial to the defendant. Dr. Lund stated that the danger lies in the fact that people "look to forensic experts, or turn to them to provide valuable information in reaching a decision," and may feel that the expert's statement of the LR "is the unique interpretation for the information presented[.]" (ECF 78, PageID.2674, 2793.) Dr. Lund stated that over the last three years, he has yet to be shown even one single instance of "a transcript of testimony or an example of a report" that shows "a presentation of the LR with some careful consideration to the influence that model choices may have." (ECF No. 78, PageID2783, 2793.) Dr. Lund also testified that he has observed a "very common tendency" that people, when presented with competing hypotheses, misinterpret it as the probabilistic characterization "about the truth of the hypothesis as opposed to the ***plausibility*** of the evidence under the hypothesis[.]" (ECF No. 78, PageID.2785-86.) He explained that unless the material has been "carefully decomposed so that a person clearly understands the distinction between the two," this mistake, also known as the prosecutor's fallacy, is a natural mistake. (ECF No. 78, PageID.2786.)

An expert's training on how to testify in court, and the use of as qualitative chart such as the one fashioned by Mr. Nye at MSP based on his own personal judgment, are also insufficient. (ECF No. 77, PageID.2698-99.) The expert cannot provide the necessary context because there are no systematic studies on the variability of LRs from a single DNA mixture. Although experts can testify on the many subjective choices that led to the LR, they cannot provide the necessary information about where the LR falls within the range of reasonable interpretations. In order for a LR to be useful to a jury, "the range of plausible other interpretations [must be] articulated and explored in some thorough manner." (ECF No. 78, PageID.2793.) Failing to give a measure of uncertainty with a LR is "inconsistent with the principles of measurement science and of transferring information from one party to another." (ECF No. 78, PageID.2762.) Because there has been no careful exploration of the possible range, no witness can provide the necessary context.

The Government may argue that it offered the defendant a chance to run the program with different variables. The Government's implication that the defendant was remiss by not proposing other hypotheses is an improper attempt to shift the burden of proof to defendant. Specifically, the burden does not rest with defendant to demonstrate that a materially different result would have occurred had any of these variables been different. Rather, it is the Government's burden to demonstrate that the LR it is attempting to introduce meets the *Daubert* standard and is reliable and helpful to the trier of fact. Incidentally, and by way of example, the Government could have re-run STRmix with conditioning profiles to try to show increased reliability of the LR. (ECF No. 78, PageID.2627, 2745-46.) However, no such evidence has been offered, and for this and the foregoing reasons, the evidence is not reliable and its probative value is outweighed by the danger of unfair prejudice.

CONCLUSION AND RELIEF REQUESTED

For the foregoing reasons, the defendant, Daniel Gissantaner, respectfully requests that this Court grant his Motion to Exclude DNA Evidence.

Respectfully submitted,

SHARON A. TUREK
Federal Public Defender

Dated: July 9, 2018

/s/ Joanna C. Kloet
JOANNA C. KLOET
Assistant Federal Public Defender
50 Louis, NW, Suite 300
Grand Rapids, MI 49503
(616) 742-7420

UNITED STATES DISTRICT COURT
WESTERN DISTRICT OF MICHIGAN
SOUTHERN DIVISION

UNITED STATES OF AMERICA,

Plaintiff,

v.

DANIEL GISSANTANER,

Defendant.

/

Case No. 1:17-cr-130
Hon. Janet T. Neff
United States District Judge

DEFENDANT'S SUPPLEMENTAL POST-HEARING BRIEF REGARDING TESTIMONY OF COURT-APPOINTED EXPERTS

The defendant, Daniel Gissantaner, by and through his attorney, Assistant Federal Public Defender Joanna C. Kloet, respectfully offers the following 15 points from the experts' testimony.[1]

1. Mixture interpretation is extremely subjective, leading to great variability in the likelihood ratio (LR).

DNA mixture interpretation "is one of the greatest challenges to forensic DNA typing" and may be influenced by several biological phenomena. (PageID.3356; PageID.4172-79.) Determining the number of contributors (NOC) is not an exact science. (PageID.4028.) As materials used in analysis increase in sensitivity, marginal samples are "getting harder and harder to interpret." (PageID.3362, 4126, 4182.) Interpretation methods have changed substantially in just the past few years and proper training is critically important. (PageID.4075-76.)

2. The MSP policy provides little guidance on when use of STRmix is appropriate.

A LR generated by STRmix is only reliable if it is properly applied, and experts caution against the temptation to submit all complex mixtures to software. (PageID.4000, 4189-90.) Here,

[1] Counsel cites only docket page references and admitted exhibits. The majority of citations reference the testimony from the July 8, 2019 hearing, but where necessary, other portions of the record also are cited.

the MSP policy manual contains "very little guidance" regarding the appropriateness of using STRmix, and has no guidance on certain points. (PageID.4167, 4190.) Further, the MSP analyst ignored even this minimal guidance, disregarding both MSP's saturation threshold and its directive to re-run the sample. (PageID.4113-16.) As a result, she discarded evidence of four contributors and concluded that the NOC was three. (PageID.4117, 4029, 4186-87.) With four contributors, the sample would have fallen outside MSP's limits for the use of STRmix.[2] (PageID.4187.)

3. The analyst lacked important background information.

A true LR does not exist for a crime scene sample if the contributors are unknown. (PageID.4126.) However, background information can increase confidence regarding the presence of a given contributor. (PageID.3359, 4090-93.) Here, information that three individuals (Cory Patton and two police officers) apparently contacted the firearm was not provided to the analyst. This information could have assisted in the interpretation of the sample, including the determination of the NOC. Without it, the LR is not reliable.

4. The technology used by MSP is obsolete.

As the government points out, science persistently improves. (PageID.4205.) Indeed, ESR has commercially released at least 10 versions of STRmix since the version 2.3.07 used by MSP in this case. (PageID.4066-68, Deft. Exh. SS.) Upgrades materially changed the software, including the ability to account for forward stutter and vary the NOC estimate. (PageID.4031-32, 4051, 4069, 4090, 4117, 4175.)[3] These improvements would have led to a more robust STRmix analysis and provided an opportunity to compare a range of LR results.[4] (PageID.2793.)

[2] Dr. Coble testified that one way to determine whether the program behaved properly is to review the output, specifically referencing the Gelman-Rubin statistic as one measure. (PageID.4050.) Here, the statistic exceeded the acceptable level. (PageID.2619.)

[3] The kits and software used by MSP in this case also have undergone revisions and updates. (PageID.4077, 4161-62.)

[4] The Canadian municipality study, offered by the Government, utilized versions 2.4 and 2.5, more recent than version 2.3.07. (Gov. Exh. 37, p 8.)

5. MSP's significantly high contamination rate affects accurate interpretation.

MSP's drop-in frequency is the rate of contamination observed in the lab during internal validation. (PageID.4118-20.) In theory, if that frequency were set at .99, then 99% of the time the lab is seeing drop-in, affecting reliability. (PageID.4123.) Dr. Coble conceded that he does not personally know of a higher drop-in rate at any other lab. (PageID.4125.) This extraordinarily high rate not only affects the analyst's interpretation, but also gives STRmix considerable "latitude" to decide whether a given peak is true biological material. (PageID.4175-76.)

6. MSP did not independently determine the boundaries for use of STRmix.

The purpose of an individual lab's internal validation is to determine its boundaries for use. (PageID.4070.) MSP did not itself determine the limits for the use of STRmix. (PageID.4190.) Rather, it relied on others' conclusions. (PageID.4168.) Additionally, MSP's validation study results appear to be "at odds with what has been in the published literature." (PageID.4165-66.)

7. MSP's internal validation study did not provide sufficient information to determine the reliability of the program on low-level samples in the lab.

MSP did not set forth the actual LRs generated by the program on low-level samples, or repeat the tests on those samples. (PageID.4116-47, 4163.) In fact, the results of its validation of low-level samples appear "inconsistent with what's been published in the literature about the lower bounds of what you need to be able to get a reliable result." (PageID.4184.)

8. STRmix falsely includes a non-contributor at an intolerably high rate.

Efforts intended "to minimize the chance of false-negatives correspondingly increase the risk of false-positives." (PageID.4136.) In a recent study, even with newer versions of STRmix, in nearly 1 out of 100 cases, a person was wrongfully included in the mixture. (PageID.4026, 4099; Gov. Exh. 37.)

9. STRmix creators did not meaningfully involve the expertise of independent software professionals in its development.

Software programs contain code, and changing or modifying a program's features involve writing or changing its code. (PageID.4106-07.) Other "[s]oftware failures have led to established practices for verification and validation (V&V) of software." (ECF No.140, PageID.3404-05.) However, STRmix does not comport with these well-known industry standards. (PageID.4155.)

10. Virtually no independent publications have validated STRmix.

Developmental validations of STRmix are not independent because developers have a vested interest in the program appearing trustworthy and accurate. Furthermore, internal validations are not peer-reviewed publications, and individual labs are incentivized to report that the program functions properly because they hope it will solve previously "unsolvable" questions. (PageID.4171-73, 4190-91.)

11. The proprietary nature of STRmix has prevented critical review of the software.

STRmix is not an open-source software program, thereby inhibiting comprehensive inspection by the community. (PageID.4093, 4111, 4185.) An extensive review of STRmix source code, which would require thousands of man-hours, has never been completed. (PageID.4185-86.)

12. The relevant community has not had adequate opportunity to review and compare STRmix with other programs.

ESR's promotional efforts, directly targeted at the United States government, suggest that its campaign prevented a meaningful opportunity for evaluation of other programs. (PageID.4062-4065.) At least 13 or 14 different PGS programs exist, and most use a completely different type of modeling from STRmix. (PageID.4086, 4096.) Fully continuous models like STRmix appear to always produce higher LRs than semi-continuous programs and always support the prosecution.

(PageID.4097-98; Gov. Exh. 38, p 149.) Thus, fully continuous models must be used "with extreme caution" on low template DNA samples. (PageID.4169; Gov. Exh. 38, p 149.)

13. Studies demonstrate significant differences between LR generated in repeat tests.

LR results can vary by several orders of magnitude in a single program. (PageID.4024, 4094-96, Gov. Exh. 38, p 145.) Without knowing the range of outputs generated within and between software programs, a juror is unable to attribute appropriate meaning to the LR. (PageID.2764.)

14. MSP's deployment of STRmix was not meaningfully audited.

MSP's use of STRmix here, as well as MSP's audit, was conducted without the benefit of ***any*** formal standards. (PageID.3353, 4015, 4045, 4102.) AAFS and OSAC have not issued final standards for PGS. (PageID.4015.) The ISFG's nonbinding recommendations for validation were not published until after STRmix was validated and applied here. (PageID.4059-60, 4065-66.)

15. The use of STRmix in this case exceeds the recommended limits.

The PCAST report—issued after MSP completed its analysis in this case—cautioned against using STRmix on samples where the minor contributor is responsible for less than 20% of the mixture. (PageID.4034, 4066.) Here, Mr. Gissantaner's purported sample is purportedly 7%, and represents less than a tenth of the DNA that would be left by a single fingerprint. (PageID.4184.)

Respectfully submitted,

SHARON A. TUREK
Federal Public Defender

Dated: July 30, 2019

/s/ Joanna C. Kloet
JOANNA C. KLOET
Assistant Federal Public Defender
50 Louis, NW, Suite 300
Grand Rapids, Michigan 49503
(616) 742-7420