LEGAL RESOURCE COMMITTEE MEMORANDUM[†]

QUESTION ON HYPOTHESIS TESTING IN ASTM 2926-13 AND THE LEGAL PRINCIPLE THAT FALSE CONVICTIONS ARE WORSE THAN FALSE ACQUITTALS

QUESTION POSED

On February 2, 2016, Forensic Science Standards Board (FSSB) member José Almirall asked the Legal Resource Committee the following question:

> I know that a basic tenet in the American legal system is that we would rather let free a guilty person than incarcerate an innocent one, if we had a choice between the two evils. Is there a mandate that we (scientific expert witnesses) behave in a certain way to ensure this?
>
> The reason I ask is that my scientific colleagues . . . are proposing that we set up our statistical tests to favor making one error (allowing a guilty defendant go free) over the possibility of making the error that we will wrongly convict.
>
> My argument is that scientists don't get this far in the *decision* process and that guilt and innocence is the purview of the jury [and] judge . . . . Our job as scientists is to present the evidence, and either error is equally bad from the scientific viewpoint (this is my opinion). . . . Does my logic make legal sense?

In response to a request for an example, he added on February 21, that:

> The Forensic Science Standards Board is considering a recommendation of the Chemistry Scientific Area Committee to place ASTM E2926 on NIST's Registry of Approved Standards. A member of the Chemistry SAC wrote the following:
>
> > The statistical matching process currently in the standard needs further examination to determine if there are more appropriate statistical criteria that could be used. As implemented now, the matching process takes a default position that every comparison will result in a match unless there is sufficient evidence in the data to prove otherwise. A consequence of this approach is that as the uncertainty of the measurement method used for comparison increases, the false match rate increases. This means that the effect of measurement uncertainty in this procedure does not work in favor of the defendant, a principle that is central to our justice sys-

[†] Memorandum from the Legal Res. Comm., to the Org. of Sci. Area Comms. for Forensic Sci., Nat'l Inst. of Standards & Tech. (rev. ed. Oct. 7, 2016). The *Harvard Law Review Forum* is reprinting this memorandum to accompany the Forensic Commentary Series. The memorandum has been only lightly edited for typographic formatting, and citations may not necessarily conform to the *Bluebook*. Original pagination is indicated symbolically.

tem. Two potential alternative statistical approaches to the current matching process include statistical $\perp_2$equivalence testing, as approved by the FDA for establishing a match between the biological activity of a generic drug and an existing approved drug, and likelihood ratios (or Bayes factors), as used in forensic DNA analysis. Both of these types of procedures give the benefit of the doubt caused by measurement uncertainty to the defendant and effectively operate so that the chance of a false match decreases or stays constant as measurement uncertainty increases.

INTRODUCTION[1]

ASTM 2926-13 (Standard Test Method for Forensic Comparison of Glass Using Micro X-ray Fluorescence (μ-XRF) Spectrometry) includes statistical procedures for determining whether two specimens of glass are "distinguishable" (and hence "did not originate from the same source of broken glass") or instead are "indistinguishable" (and hence possibly "originated from the same source of glass").[2] The part of the Standard Test Method on "interpretation of comparisons" (§ 10.7) is reproduced as an appendix to this memorandum. We have previously urged that the Standard be revised in several ways, including offering more explicit guidance on how analysts should describe the analytical results and their interpretations in their reports and testimony. This memorandum does not change our conclusion that important revisions should be made before adding this Standard to the OSAC Registry, and it is not intended to depart from our previous comments.

As we understand the question, you would like to know whether the procedures in § 10.7 are consistent with legal principles for allocating the risk of error in criminal cases. *We believe that the principle that false convictions are more serious than false acquittals does not preclude the use of these procedures*, but neither does it prevent statistical equivalence testing or the use of a likelihood ratio or Bayes factor to inform the judge or jury of the probative value of the spectroscopic results. None of these methods is foreclosed solely by the need to guard against false convictions at the expense of an increased proportion of false acquittals.

We do not address other questions, such as ascertaining (1) which approach would best meet the needs and goals of the legal system, (2) the extent to which the scientific literature establishes that the process for classifying samples is reliable and has known conditional error

[1] The committee adopted this statement by a unanimous vote. Those voting were Lynn Garcia, Christine Funk, Jennifer Friedman, Ted Hunt, David Kaye, David Moran, Christopher Plourd, Barry Scheck, and Ronald Reinstein.

[2] ASTM 2926-13 Introduction.

rates, and (3) the suitability of the match/no-match criteria of the test method for the purpose of investigation as opposed to proof at trial.

ANALYSIS AND EXPLANATION

⊥3 To explain the basis for the above conclusions, we offer the following observations about the law of evidence as it applies to statistical evidence:

(1) *The law prizes neutral experts.* Ideally, the law of evidence contemplates that a scientist providing expert evidence inform the judge or jury of facts or opinions that would assist it in resolving disputed questions of fact without favoring one party over the other. Although scientific "truth" may be elusive, it is the goal in civil and criminal cases alike. Whether appearing at the request of one party or appointed by the court, "statisticians and other experts should take the role of impartial educator . . . ."[3]

(2) *The law permits experts to present statistical findings using a variety of methods of statistical inference.* Broadly speaking, the three major statistical approaches to evaluating hypotheses in the light of data (or expressing the uncertainty in these hypotheses) are frequentist, likelihood, and Bayesian.[4] Frequentist hypothesis tests, which are the subject of your question, are routinely used in litigation, but Bayesian analyses also have been held to be admissible.[5]

(3) *The legal burden of persuasion pertains to a level of subjective confidence that is reasonably based on the evidence.* In *Santosky v. Kramer*, 455 U.S. 745 (1982), the Supreme Court, quoting from earlier cases, observed that:

> The function of a standard of proof, as that concept is embodied in the Due Process Clause and in the realm of factfinding, is to "instruct the factfinder concerning the degree of confidence our society thinks he should have in the correctness of factual conclusions for a particular type of adjudication."[6]

In most civil cases, a plaintiff need only prove that a set of facts that warrant a verdict of liability are probably true. In criminal cases, the burden of persuasion is the much higher standard of proof beyond

---

[3] PANEL ON STATISTICAL ASSESSMENTS AS EVIDENCE IN THE COURTS, NATIONAL RESEARCH COUNCIL, THE EVOLVING ROLE OF STATISTICAL ASSESSMENTS AS EVIDENCE IN THE COURTS 159 (Stephen E. Fienberg ed., 1989). *See also* DAVID H. KAYE ET AL., THE NEW WIGMORE ON EVIDENCE: EXPERT EVIDENCE (2d ed. 2011).

[4] *See, e.g.*, VIC BARNETTE, COMPARATIVE STATISTICAL INFERENCE (3d ed. 1999); FRANCO TARONI ET AL., DATA ANALYSIS IN FORENSIC SCIENCE: A BAYESIAN DECISION PERSPECTIVE (2010).

[5] *E.g.*, Plemel v. Walter, 735 P.2d 1209 (Or. 1987) (parentage testing); KAYE ET AL., *supra* note 3, at §§ 14.2.2, 14.3.2 (criticizing some of the parentage testing cases for allowing an expert to select a prior probability).

[6] 455 U.S. at 754–55.

a reasonable doubt. The Court has written that the civil "preponderance of the evidence" standard implies that the parties "share the risk of error in roughly equal fashion, whereas "[i]n the administration of criminal justice, our society imposes almost the entire risk of error upon itself."[7] ⊥4This understanding of the state's burden in criminal cases reflects the aversion to false convictions that has been expressed time and again in many cultures.[8]

(4) *The "confidence" related to the burdens of persuasion is distinct from the "confidence" associated with frequentist "confidence intervals" and "significance levels."* For testing statistical hypotheses, frequentist methods look to the long-term properties of decision rules. Neyman-Pearson hypothesis testing involves a "null hypothesis" (H_0) and an "alternative hypothesis" (H_1) and rejects the former in favor of the latter if and only if the value of a test statistic falls within a "rejection region" chosen to keep the probability of a false rejection (a "Type I error") at or below a specified "significance level" (α). For a 0.05 level, for example, a decisionmaker will reject the null hypothesis approximately 1 time out of 20 when it expresses the true state of affairs. Thus, in the worst-case scenario for Type I errors (in which H_0 is true in every case), the decisionmaker will err in rejecting H_0 about 5% of time.[9]

ASTM 2926-13 § 10.7 refers to "the confidence level of an association" and lists a "±3s" rule that "corresponds to 99.7% of a normally distributed population." Using a confidence interval with coverage probability $1 - \alpha$ to decide whether a difference between measurements is significant is equivalent to testing H_0 at a significance level α.

---

[7] Addington v. Texas, 441 U.S. 418, 423–24 (1979). Some commentators maintain that these standards describe a posterior probability conditioned on all the evidence in the case and follow from the relative utilities of the two types of mistaken verdicts in the framework of Bayesian decision theory. *E.g.*, John Kaplan, *Decision Theory and the Factfinding Process*, 20 STAN. L. REV. 1065 (1968); D.H. Kaye, *Clarifying the Burden of Persuasion: What Bayesian Decision Rules Do and Do Not Do*, 3 INT'L J. EVIDENCE & PROOF 1 (1999). Because the disutility of a false criminal conviction greatly exceeds that of a false acquittal, a rule that requires a large probability of guilt is necessary to maximize expected utility.

[8] *See* Alexander Volokh, n *Guilty Men*, 146 U. PA. L. REV. 173 (1997).

[9] As noted in point (1), the law asks that scientists present findings without favoring one side or the other. But this does not mean that scientists must construct hypothesis tests that have equal Type I and Type II error probabilities. If it did, the limits of detection based on signal-to-noise-ratios of 10 in the Standard would be too demanding. One might imagine that in civil cases, a rejection region should be defined so as to equalize Type I and Type II error probabilities, but this notion does not implement the more-probable-than-not burden of persuasion for civil cases. David H. Kaye, *Hypothesis Testing in the Courtroom*, *in* CONTRIBUTIONS TO THE THEORY AND APPLICATION OF STATISTICS 331 (Alan Gelfand ed., 1987). Equating the utilities of each type of error within a Bayesian decision-theoretic framework would produce a rejection region that implements the principle that both errors are equally serious.

Thus, the description of the matching procedure suggests that it has a Type I error probability of only 0.003.[10]

The confidence coefficient $1 - \alpha$ is not the same as the subjective confidence required for a guilty verdict. The 0.003 level applies, at best, to the decision as to each element that is detected, and not to the analyst's decision that glass fragments are (or are not) distinguishable in all their physical and chemical properties. Furthermore, as the remarks from the SAC statistician indicate, the statistical test privileges the statistical hypothesis that the element's true concentration is the same in both samples — which is at the opposite pole from the legally relevant null hypothesis that the questioned and the known samples come from different specimens (indicating the defendant's innocence).

(5) *The law does not dictate the choice of a significance level for a test of the statistical null hypotheses (same-composition).* The distinction between the statistical hypothesis that two samples have the same elemental composition and the legal hypothesis that they came from the same source does not necessarily render the desired significance level legally incorrect. Because the significance level or its complement cannot be equated to the degree of certainty required for a conviction based on all the evidence, the latter does not clearly inform the choice of the significance level.

Conventionally, significance levels such as 0.05 or 0.01 in biomedical and social sciences are used to protect against false claims of new discoveries. Particle physicists are even more demanding, sometimes requiring a 0.0000001 level before announcing the discovery of a new particle. In these contexts, false alarms are seen as worse than false misses. So the position that there is no true difference, no true effect, or no true discovery — the null hypothesis — is the default position, and it takes strong evidence to move the science off this baseline. In sum, the choice of a particular significance level is essentially a science-policy decision, and it is not uncommon for forensic scientists to urge greater caution to avoid false inclusions than false exclusions.[11]

---

[10] Whether the description in § 10.7 is properly phrased and is statistically correct is beyond the scope of this memorandum. Such matters could be addressed by the FSSB's Statistics Task Group.

[11] *E.g.*, E.J. Garvin & R.D. Koons, *Evaluation of Match Criteria Used for the Comparison of Refractive Index of Glass Fragments*, 56 J. FORENSIC SCI. 491, 499 (2011); R.D. Koons & J. Buscaglia, *Interpretation of Glass Composition Measurements: The Effects of Match Criteria on Discrimination Capability*, 47 J. FORENSIC SCI. 505, 512 (2002); B.D. Gaudette, *The Forensic Aspects of Forensic Textile Examination*, *in* 2 FORENSIC SCIENCE HANDBOOK 209, 255 (R. Saferstein ed., 1998). *But see* Ian W. Evett, *Expert Evidence and Forensic Misconceptions of the Nature of Exact Science*, 36 SCI. & JUST. 118, 120 (1996) (arguing against "a widespread view that the expert must quote an assessment of the evidence which is conservative in the sense of erring so as to favour the defendant").

However, the prosecution is not the only party that may offer evidence apparently linking an individual to a crime. A defendant in a criminal case may seek to establish that a different individual is associated with the crime scene, and the same evidence can arise in civil cases.

Courts expect scientists to provide evidence using methods and analyses that are no less rigorous than the norm for scientific inquiry and publication.[12] In toxic tort and employment discrimination litigation, they generally insist on statistically significant results at conventional levels.[13] Certainly, if a decision procedure has been shown to have good operating characteristics — if it has small Type I *and* Type II error probabilities — then it should be both scientifically and legally acceptable.

(6) *A liberal matching rule can be legally acceptable.* On its face, ASTM 2926-13 has a broad matching rule, and making "the same elemental composition" the null hypothesis has the seemingly perverse effect that you quoted: "as the uncertainty of the measurement method used for comparison increases, the false match rate increases." The effect is the consequence of the incongruity between the *legal* hypothesis of different-source (consistent with innocence) and the *statistical* null hypothesis of equal-elemental-composition (consistent with same-source and guilt).

Although this incongruity has very serious implications for how a finding of "no significant difference," "indistinguishable," "matching," or "consistent" should be presented in court, it does not necessarily render the finding inadmissible. The major legal demand on a scientifically validated and reliable process for making measurements and inferences is that the conclusion be reported with a suitable description of its probative value. Within the match/no-match framework of ASTM E2926-13, this means that reports and testimony must address what is often called the "second stage"[14] in the analysis of identification evidence — namely, estimating the probability of a match for questioned specimens originating from glass elsewhere in the population. If this is done, the fact that it is harder for a laboratory whose measurements are imprecise to reject H_o is balanced by the fact that the laboratory must report a less impressive estimate of the probability of a false inclusion. Its match window is wider, so there will be more matching glass in the population, and this is reflected in the second-

---

[12] *See, e.g.*, KAYE AT AL., *supra* note 3 (discussing case law following Daubert v. Merrell Dow Pharms., 509 U.S. 579 (1993)).

[13] *Id.*

[14] *E.g.*, Ian W. Evett, *The Theory of Interpreting Scientific Transfer Evidence*, *in* 4 FORENSIC SCI. PROGRESS 143, 146, 148 (A. Maehly & R.L. Williams eds., 1990); E.J. Garvin & R.D. Koons, *Evaluation of Match Criteria Used for the Comparison of Refractive Index of Glass Fragments*, 56 J. FORENSIC SCI. 491, 491 (2011).

stage statistic that indicates the probative force of the match.[15] The courts confronted this same issue in holding admissible DNA matches ⊥7based on wide match windows used to define VNTR alleles.[16] "Match-binning" was deemed legally acceptable on the theory that it was shown that the match criteria rarely failed to include truly matching fragments, and the random-match probabilities computed with the corresponding bin frequencies could be presented to jurors to enable them to assess the legal significance of a finding that the RFLP fragments were indistinguishable under the statistical rule for matching.[17]

* * *

In sum, the widespread aversion to false convictions does not make reporting the result of a frequentist statistical test with small Type I and Type II error probabilities inadmissible per se in court. Even though the choice of equal elemental composition for a null hypothesis makes the application of the usual statistical terminology of Type I and Type II errors potentially confusing, it is not in itself grounds to exclude evidence of a match. The selection of the particular threshold that separates matches from nonmatches with the null hypothesis in ASTM 2926-13 reflects a scientific policy judgment. We are not in a position to opine on whether the judgment in this instance departs from normal scientific practice, which would render the test method inadmissible, but we note that a report of a match without more information about the probability of a match to other glass in the relevant population would not fulfill the expert's role of impartially and adequately educating the trier of fact about what the scientific mea-

---

[15] David H. Kaye, *The Relevance of "Matching" DNA: Is the Window Half Open or Half Shut?*, 85 J. CRIM. L. & CRIMINOLOGY 676 (1995). If no valid random-match probability can be computed, admissibility of a match is debatable. Unless it is somehow made clear to the judge or jury that an unknown number of people other than the defendant might be similarly associated with the crime scene, there is a serious danger that the jury will overvalue the test finding. Conclusions with no reference to the probability of a match in relevant population can create real misunderstandings — triers of fact may afford more weight to the association than they should, thinking that "well it must have been him because how else did that 'matching' or 'indistinguishable' glass get there?" The Standard should help forensic scientists clearly express the strength and limits of their analysis in a way that neither underrepresents nor overinflates the connection between a known and a questioned piece of glass.

[16] DAVID H. KAYE, THE DOUBLE HELIX AND THE LAW OF EVIDENCE (2010).

[17] *Id.* There are cases holding DNA match testimony inadmissible without an estimate of the probability of a match in the relevant population. *E.g.*, Commonwealth v. Cole, 41 N.E.3d 1073, 1083–84 (Mass. 2015); State v. Tester, 968 A.2d 895, 909 (Vt. 2009). Other courts do not insist on a quantified probability, but still deem the evidence "inadmissible absent some accompanying interpretive evidence regarding the likelihood of the potential match." People v. Coy, 620 N.W.2d 888, 898 (Mich. Ct. App. 2000). A smaller number of jurisdictions admit a statement that a defendant could not be excluded even with no further explanation. *E.g.*, Rodriguez v. State, 273 P.3d 845, 851 (Nev. 2012).

surements establish. The OSAC is addressing these limitations by developing interpretation standards to better inform the trier of fact about the significance of glass evidence.

⊥8

APPENDIX

Section 10.7 of ASTM 2926-13 Standard Test Method for Forensic Comparison of Glass Using Micro X-ray Fluorescence (μ-XRF) Spectrometry

10.7 *Interpretation of Comparisons:*

10.7.1 *Peak Identification* — Reproducible differences in detected elements between specimens demonstrate that the specimens have different sources. When peak identification does not discriminate between the specimens, further spectral comparisons should be conducted.

10.7.2 *Spectral Comparisons* — Reproducible differences in spectral shapes and relative peak heights between specimens may indicate that the specimens have different sources. Peak intensity ratios can be calculated to demonstrate this difference. When evaluation of spectral shapes and relative peak heights do not discriminate between the specimens, peak intensity ratios should be calculated.

10.7.3 *Peak Intensity Ratio Comparisons* — Reproducible differences between specimens in peak intensity ratios can demonstrate that the specimens have different sources. One of the two following statistical measures is recommended to assess the association or discrimination of the samples based on elemental ratios:

10.7.3.1 *Elemental Ratio Range Overlap* — For each elemental ratio, compare the range of the questioned specimen replicates to the range for the known specimen replicates. Because standard deviations are not calculated, this statistical measure does not directly address the confidence level of an association. If the ranges of one or more elements in the questioned and known specimens do not overlap, it may be concluded that the specimens are not from the same source.

10.7.3.2 *±3s* — For each elemental ratio, compare the average ratio for the questioned specimen to the average ratio for the known specimens ±3s. This range corresponds to 99.7% of a normally distributed population. If, for one or more elements, the average ratio in the questioned specimen does not fall within the average ratio for the known specimens ±3s, it may be concluded that the samples are not from the same source.