REFLECTIONS ON GLASS STANDARDS: STATISTICAL TESTS AND LEGAL HYPOTHESES

David H. Kaye

The Pennsylvania State University, State College, PA USA

***Abstract**: The past 50 years have seen an abundance of statistical thinking on interpreting measurements of chemical and physical properties of glass fragments that might be associated with crime scenes. Yet, the most prominent standards for evaluating the degree of association between specimens of glass recovered from suspects and crime scenes have not benefitted from much of this work. Being confined to a binary match/no-match framework, they do not acknowledge the possibility of expressing the degree to which the data support competing hypotheses. And even within the limited match/no-match framework, they focus on the single step of deciding whether samples can be distinguished from one another and say little about the second stage of the matching paradigm–characterizing the probative value of a match. This article urges the extension of forensic-science standards to at least offer guidance for criminalists on the second stage of frequentist thinking. Toward that end, it clarifies some possible sources of confusion over statistical terminology such as "Type I" and "Type II" error in this area, and it argues that the legal requirement of proof beyond a reasonable doubt does not inform the significance level for tests of whether pairs of glass fragments have identical chemical or physical properties.*

1. INTRODUCTION

Much of the writing in this special issue of *Statistica Applicata – Italian Journal of Applied Statistics* proceeds from a Bayesian perspective. And for good reason. The Bayes' factor, likelihood ratios, and prior and posterior probabilities can be immensely helpful in reasoning about the weight and implications of forensic-science evidence (e.g., Aitken and Taroni, 2004; Aitken et al., 2010; Kaye 2016). Nevertheless, some fields of forensic science remain dominated by categorical conclusions of the kind that emerge from classical hypothesis testing. Even DNA testing, in which the movement toward a weight-of-evidence approach has been most influential (Balding and Steele, 2015; Buckleton, Bright, & Taylor, 2016; Carrecedo, 2015; Evett and Weir, 1998), typically uses "analytical thresholds" and "stochastic thresholds" based on predetermined signal-to-noise ratios for deciding whether a measurement establishes the presence of an allele. Likewise, to ascertain whether glass fragments originated from a particular sheet of glass, forensic chemists measure physical properties and chemical composition of glass fragments, then assess them with up-or-down tests of various degrees of statistical sophistication (Almirall, 2013).

The most prominent standards for making and analyzing measurements on glass come from ASTM, Inc., a private organization that develops standards for everything from making nuts and bolts to measuring lead levels in gasoline. The ASTM standards for forensic glass analysis have been praised as the outcome of a "rigorous standard development process" and as ready for adoption by the U.S.-government-supported Organization of Scientific Area Committees for Forensic Science (Almirall and Trejos, 2016, 230). These standards may contribute greatly to the proper use of powerful and precise instruments for chemical and physical analysis of glass, but their adequacy at the interface of law and statistics is less obvious.

This essay surveys some of the statistical aspects of the simplest of these glass standards. It argues that to the extent that criminalists' findings should include categorical judgments, the standards should specify the statistical procedures that perform best for classifying two specimens of glass as having come from the same source; they should offer guidance on how to describe the probative value of these statistically-based decisions; and they should acknowledge the existence of alternative modes of statistical analysis. Although the argument breaks no new statistical or legal ground, it serves to clarify the relationship between legal principles such as the presumption of innocence and proof beyond a reasonable doubt, on the one hand, and the choice of a null hypothesis and a significance level in a classical hypothesis test on the other.

2. REFRACTIVE INDEX AND ASTM E1967-11A

The index of refraction (RI, or η) of a substance indicates how quickly light travels in that medium (v) compared to its speed c in a vacuum ($v = c/\eta$) and how its wavelength λ in that medium differs from its wavelength λ_0 in a vacuum ($\lambda = \lambda_0/\eta$, where η depends on the wavelength). Various methods for measuring RI exist, and ASTM E1967-11a describes a "Standard Test Method for the Automated Determination of Refractive Index of Glass Samples Using the Oil Immersion Method and a Phase Contrast Microscope." This standard has no distinct section on statistical methods for the interpretation of any measurements. At one point, it states that "precision and bias of this test method should be established in each laboratory that employs it. Confidence intervals or a similar statistical quality statement should be quoted along with any reported [RI] value. For instance, a laboratory may report that the error for the measurement, using a reference optical glass is 0.00003 units" (ASTM E1967-11a § 3.5). It adds that "[p]recision of refractive index measurements should meet the original equipment manufacturers specifications" (ibid. § 7.1) and that "[s]ince the measurement of the sample [RI] is a direct comparison to the standard reference glasses used, no bias exists [in intralaboratory comparisons]" (ibid. § 7.2).

To readers who are not analytical chemists (and perhaps to some who are), this advice might seem cryptic. What is "0.00003 units"? RI, being a ratio of speeds, is a dimensionless quantity.[1] If "the error for the measurement" is the actual measurement minus the true value (the usual definition), then the error for a particular measurement of the RI of glass from the crime scene or suspect is unknown – if we knew the true value of that glass, we would not have needed to make the measurement.[2]

Perhaps what the standard calls "the error for the measurement" is the "precision of refractive index measurements." Precision usually is measured with a statistic such as the standard error in replicate testing, and manufacturers quote 0.00003 for the "standard deviation" or "precision" of their instruments (Foster + Freeman Ltd. 2016; Microtrace 2016; SWGMAT 2004). This standard error might be the basis for the confidence interval that "should be quoted along with any reported [RI] value." But what confidence level should be used, and how should the confidence interval or other "statistical quality statement" be presented? If the interval is narrow, should the expert describe the measured value as being of high "statistical quality"?

Not only is the phrase "error for the measurement" and its illustrative number 0.00003 somewhat unclear, but what matters is not the variability of a single measurement of the refractive index η but rather the *difference d* between the estimated value n_q of the RI of the "questioned" sample and the estimated value n_k of the RI of "known" sample ($d = n_q - n_k$). Hence, there are two sources of variability to consider, and the confidence interval for an estimate of η_k alone understates the relevant uncertainty.

In practice, estimates of η_k and η_q would come from multiple measurements on the questioned fragment and the known glass. Moreover, several physical samples might be drawn from each specimen to assess the variability in η at different points in the glass. Indeed, it is thought "the precision of the method is typically better than the measurable variation of a glass object" (SWGMAT 2004,

[1] Consequently, it does not come in "units," let alone the standard units mentioned in sections 1.4 and 3.5 of ASTM E1967-11a.

[2] Of course, for a measurement on a certified reference sample (the "reference optical glass" whose RI is known to a desired number of decimal places), the error will be known (Joint Committee for Guides in Metrology, 2008, § 2.16(a)). But that is a different measurement than those of interest in the case.

§ 7.2.3.1). Plainly, a statistically and forensically satisfactory standard needs to address much more than the precision, bias, and confidence intervals for the single measurements mentioned in ASTM 1967-11a. Work in this direction is underway (Almirall, 2016).

3. RI MATCHING AS A FORM OF HYPOTHESIS TESTING IN COURT

A case involving both sets of glass measurements is *Johnson v. State,* 521 So.2d 1006 (Ala. Ct. Crim. App. 1986).[3] A man in a small town in Alabama who sold jewelry from his home was killed in an exchange of gunfire with two robbers. There was evidence that he had shot one of the robbers, and the defendant had a bullet lodged in his back. A toolmark analysis of the surgically removed bullet "did not produce a definite determination that [the defendant's] revolver actually fired the bullet." (Ibid., 1009). However, a bullet had gone through a "pane on the back door," the bullet taken out of the defendant's back had "glass imbedded in its nose," and none of the bullets recovered from the house had glass on them. (Ibid.) To see whether the bullet lodged in defendant's back might be the one that went through the glass, the FBI compared the fragment (the "questioned" specimen) to a "known" specimen of the glass in the back door. The court's opinion does not report any measured values,[4] but it states that the FBI found that the specimens "matched, with no measurable discrepancies" in their RIs. (Ibid.)

A more careful statement would be that the measurements for the two specimens were close or similar–meaning that the observed values were within some arithmetic difference $d = n_q - n_k$ of one another, and that if they had originated from the same location on the pane, a difference more extreme (farther from 0) would occur some proportion p of the time. An experiment might show that the errors in replicate measurements of a standard reference are normally distributed with mean 0 and a standard deviation σ. An estimate for σ from the experiment might be 0.00003 or some other numbers, and this value might be used in conjunction with the normal error model to estimate p, but standing alone, s offers little guidance.

[3] This description of the RI analysis in the case may be an oversimplification. The laboratory may have made more than one measurement for the two glass samples and compared the two sets of values in some manner.

[4] For simplicity, I will assume that only one measurement on each fragment was made. The essential ideas would be the same for a comparison of multiple measurements of each specimen.

In the significance testing framework, p is a p-value for the null hypothesis $\delta = \eta_q - \eta_k = 0$, where η is the true RI. Suppose p were 0.03 that is, only 3% of the time would the differences between the measured values be expected to be as extreme as d when measuring known and questioned fragments that have the same true RI. A criminalist might be tempted to reject H_0 in favor of some alternative hypothesis H_1. But we must be very careful in thinking about the relationship between the hypotheses and statements about the origin of the samples. H_0 only states that the two specimens have the same true, unknown RI—and hence might have come from the same piece of glass. H_1 only states that they have different true RIs—and hence either came from different pieces of glass or, if the known pane is spatially heterogeneous in its RI, from a different place on the known pane.

H_0, which is privileged by demanding a small p-value to overthrow it, undercuts a defense claim that the questioned glass came from somewhere else. Choosing a rejection region that makes it difficult to reject H_0 thus might seem to stack the deck against the defendant. It also means that the less precise the laboratory is in its measurements, the easier it is to report that that the specimens "matched, with no measurable discrepancies." I will return to these concerns in Section 5.

Also critical to a fair presentation of the results is the recognition that a failure to reject H_0 does not mean that the two specimens are from the same source of glass. It only means that there is insufficient evidence to conclude at a desired level of "significance" that two specimens have different RIs. The hypothesis that is of more direct interest to the legal system is that the glass associated with the defendant came from the crime scene (cf. Kaye, 2005, 95–96). I will call this the *legal same-source* hypothesis to distinguish it from the *statistical same-value* hypothesis. Its assessment requires data on the distribution of RI in the population of glass from which the questioned specimen could have come. Plainly, if all glass had the same RI, the inclusionary finding would have *no* probative value for this legally relevant question no matter what significance level was attained for that hypothesis of equal true RIs. Thus, in concluding "that the evidence was more than sufficient to allow the jury to reasonably conclude that the evidence excluded every reasonable hypothesis except that of guilt" (Ibid., 1013), the *Johnson* court also explained that "[b]ased upon F.B.I. statistical information, it was determined that only 3.8 out of 100 samples could have the same physical properties, based upon the refractive index test alone, which was performed." (Ibid., 1009).

These clarifications do not render the ASTM standard wrong or deficient in its description of how to measure RI. But they do show that the standard is incomplete in helping criminalists report on the outcome of tests in even the simplest of all possible situations – single measurements of a single variable for a single questioned and a single known specimen. ASTM E1967-11a does not acknowledge that determining that fragments of glass have “matching”, “indistinguishable”, “similar”, or “consistent” RIs is not the only (and perhaps not the best) way to convey the implications of the findings in the context of investigations and trials. And, with respect to this traditional approach, the standard makes no attempt to prescribe (or even describe) statistical criteria and methods that would be appropriate for the match/no-match decision.[5] This omission is puzzling when one considers the wealth of statistical thinking both frequentist and Bayesian over the past 50 years on statistical inference with RI and similar data. For reviews or textbooks, see Curran et al., 2000; Evett, 1990; Zadora et al. 2014.

4. RI MATCHING AS A SCREENING TEST

Of course, the *Johnson* case is 30 years old. Today, RI is not as likely to be the major factor in a criminalist’s conclusion that two glass samples share a common source. More discriminating analytical tests of chemical composition are available (Almirall and Trejos, 2016; Dorn et al., 2015; Koons and Buscaglia, 2002; Tejos et al., 2013). Consequently, it might seem that there is little need for the forensic-science standard on RI analysis to fuss with statistical niceties.

But this thought cannot withstand much scrutiny. Even if we regard RI analysis as a screening test, it matters where one sets the threshold for a positive finding. In the case of a screening test for a disease, a positive test result indicates that the disease is present. False positives can lead to patient anxiety and more invasive and expensive follow-up tests. In contrast, a positive finding for a RI comparison implies that the glass from the crime scene is *not* present, but that too has deleterious consequences. A false positive for the glass screening test not only leads the laboratory to dispense with additional tests on the glass, but it also can lead the police to falsely exclude the suspect and to fail to gather or to discount other evidence inconsistent with the exclusion. With a medical screening test, a false negative can mean that a life-saving (or less dramatic) intervention will not be pursued, whereas a false negative on the RI-screening test merely leads to additional “gold standard” testing. In these circumstances, there is a fair

[5] The ASTM references point readers to Miller’s suggestion (Miller, 1982, 165-66) of using “fixed differences in refractive index or density beyond which a conclusion of two distinct sources is made [and which was] used by many glass examiners for at least 25 years” (Almirall, 2013)—even though leaders in the field no longer recommend it (ibid.).

argument for setting the cut-point of *d* for a positive result (an exclusion) at a very high level (Bottrell, 2009; Garvin and Koons, 2011, 499).[6]

Furthermore, even within the exclusively frequentist framework, a forensic-science standard on interpreting RI measurements must clearly distinguish between the statistical hypotheses being tested and the legal hypotheses to which they relate.[7] As to the preliminary statistical hypothesis, the report can grade the force of the evidence with a *p*-value rather than an up-or-down decision about the true values of the RIs. In addition, the forensic-science standard should explain how laboratories that use the forced-decision approach can estimate the conditional error probabilities (or, equivalently, the specificity and sensitivity of their screening test).[8] Finally, a modern standard should recommend or require that reports include these statistics rather than merely state a standard error or confidence interval for a single measurement.

5. CHOOSING THE NULL HYPOTHESIS AND SIGNIFICANCE LEVEL

An overarching issue for forensic-science standards for methods that generate quantitative evidence of association between items and crime scenes using match/ no-match criteria is the selection of a reasonable significance level for a statistical hypothesis test. Ordinarily, one fixes the probability of a Type I error and seeks a test that keeps the risk of a Type II error to a minimum, where a Type I error is a false rejection of the null hypothesis, and a Type II error is a false failure to reject the null hypothesis. As we saw in Section 3, in the context of a screening or diagnostic test, a Type I error is a false positive, and a Type II error is a false negative. In signal detection theory, they would be called false alarms and missed

[6] However, there is a stronger argument for informing the investigators of how much the RI measurements support the hypothesis of a common source, if computation of that likelihood ratio is feasible. Moreover, even within the exclusively frequentist framework, to call such a cut-score "conservative," as Bottrell (2009), does, introduces a potential terminological pitfall. A high cut-score is not "conservative" in preserving the legal status quo in which a defendant is deemed innocent in the absence of compelling evidence of guilt. Instead, it "conserves" an individual's status as a suspect who might be further incriminated by more testing. (Compare the definition of "conservative" in NRC Committee (1996, 215) as "favoring the defendant.")

[7] The analyst's report – which will influence the thinking of investigators and perhaps prosecutors, defendants, judges and jurors – should make it plain that the error probabilities associated with a "match" or "nonmatch" (that is, a test of the statistical hypothesis that samples have the same value of a physical property) is not the probability of a common origin for two samples.

[8] For studies along these lines, see Dorn et al. (2015, 94-95); Garvin and Koons (2011, 242) (within-sample study of "[f]ive sheets typical of modern float glass products" for false exclusion rates with various match criteria) and the papers cited in note 10.

signals, respectively (Melsa and Cohn, 1978). In forensic science and law, one might think that they would be false inclusions and false exclusions of suspects, but the terminology is reversed (e.g., Curran et al., 2000; Dorn et al., 2015, 87; Gaudette, 1988, 255; Trejos et al., 2013, 1272). Unlike an epidemiologic study of a suspected carcinogen or a clinical trial of a new drug, the "null hypothesis" is not that there is "no association" or "no effect." It is that there is an association between the samples (insofar as they have the same RI or other properties).

Consequently, one finds statements such as "Type 2 errors are generally considered more insidious than Type 1 errors because a false association may lead to incrimination of an innocent subject" (Garvin and Koons, 2011, 499); "[a] Type II error, or false inclusion, is considered the more egregious error in forensic comparisons, because it may incriminate a truly innocent subject" (Koons and Buscaglia, 2002, 505); and "the consequences of a type II error are much more serious than the consequences of a type 1 error" because a "type II would result in wrongly incriminating evidence being presented whereas a type I error would generally result in no evidence being presented against a guilty person" (Gaudette, 1998, 255).

Yet, some ASTM glass standards countenance decisions based on three and four standard errors (ASTM 2013, § 10.7.3.2; ASTM 2012, § 10.1.4). Such expansive match criteria should give good protection against statistical same-value Type I errors, but what about the more serious legal same-source Type II

[9] For reports of error rates with a variety of matching rules within a source of fragments or across different sources, see Dorn et al. (2015) (within-source testing for RI and 10 elements of 1 pane and across-source testing of 82 sources from casework in Ontario); Koons and Buscaglia (2002, 511) (reporting that "[t]he evaluation of elemental composition data for evidentiary glass samples [209 specimens from 148 cases] shows that a very small likelihood exists for failing to discriminate between glass fragments from different sources, regardless of the matching procedure or significance level used."); Trejos et al. (2013) (interlaboratory testing of elemental composition of a small number of samples from same and different plants over varying period of time); Weis et al. (2011) (within-source, 18-element compositional analysis for one window pane and across-source analysis for 62 samples from various countries).
Two points about these studies are worth noting. First, whether the reported error rates can be used to quantify the probative value of findings in a particular case depends on the laboratory's choice of trace elements, its matching rule, and the representativeness of the tested sources with respect to the relevant population of glass for the case (e.g., Dorn et al., 2015, 94). Second, presenting Type II error probabilities from across-source studies is more appropriate than estimating Type II error probabilities with respect to the statistical hypothesis of identical RI and elemental composition (cf. Weis et al., 2011, 1276). The existence of, alas, two types of Type II errors, arises because, as noted earlier, the statistical hypothesis of equal true values for physical or chemical properties differs from the legal same-source hypothesis. If a large proportion of the glass in the relevant population would match under the laboratory's decision rule, even a very powerful statistical test (one with a very small probability for rejecting the null hypothesis of equal RI and elemental composition) would be ineffective in testing the hypothesis that the questioned samples originated from the known glass specimen.

errors?[9] Can this approach be reconciled with legal maxims and principles such as the presumption of innocence and the requirement of proof beyond a reasonable doubt? Or does the hoary legal principle that a false conviction is worse than a false acquittal (Volokh, 1997) require structuring statistical hypothesis testing that would advantage defendants over the state?

On examination, the legal maxims turn out to be more distracting than helpful here. The reasonable-doubt standard used in criminal cases applies to the totality of the evidence. When all is said and done, the prosecution has the burden of persuading the judge or jury beyond a reasonable doubt, but each individual piece of evidence should be presented for what it is worth, and not discounted because one party or the other is introducing the testimony.[10]

Even so, these general observations about the law do not shed any light on how to define the rejection region for a frequentist hypothesis test, or how high a confidence coefficient should be for drawing inferences from sample data. The legal burden of persuasion speaks directly to a posterior probability – the probability of the prosecution's hypothesis conditioned on all the evidence (e.g., Kaye, 1987). The "reasonable doubt" that has to be eliminated is doubt about the prosecution theory of the events – doubt about the state's hypothesis that the defendant committed the crime charged. In contrast, significance levels pertain to the probability of a range of evidence given a null hypothesis. They do not translate into the probability that the either the statistical or the legal null hypothesis is false, and their complement, "confidence," is not the probability that the alternative hypothesis is true (e.g., Kaye and Freedman 2011).

Of course, there is an analogy between the underlying reasons for the high burden of persuasion for the entire case in law and the reasons to use demanding significance levels for announcing new discoveries in science. Using a level of 0.05 or 0.01 in biomedical and social sciences protects against false claims of new discoveries, and journal editors do not want to be caught publishing too many such articles. With new pharmaceuticals, regulatory agencies do not want too many ineffective drugs reaching the public. Particle physicists are even more demanding, sometimes requiring a 0.0000001 level before announcing the discovery of a new particle (van Dyk, 2014, 52–54). In these contexts, false alarms

[10] When a modeling assumption is in serious doubt, however, it may be pragmatically desirable to resolve doubts so as to favor the party against whom the evidence is introduced. For example, in estimating a DNA random-match probability, one might use an extremely large value for the co-ancestry coefficient to forestall a defense argument. However, this is not really a legal issue. It is comparable to choosing a low estimate of future climate change to argue that immediate action is essential under any plausible scenario.

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. Analogously, the law regards false convictions as worse than false acquittals. Consequently, it demands a very strong case against the defendant to reject the privileged hypothesis of innocence. But the analogy does not apply at the level of an individual forensic test. In the words of one famous legal treatise, "[a] brick is not a wall" (Broun et al., 2013, 1000). Even modestly probative scientific evidence should be admissible, at least if the limitations on its probative value can be conveyed and understood.

Thus, regardless of how one determines the Type I risk for the statistical same-value hypothesis (or, for forensic-science examinations with "match" as H_0, the Type II risk) to tolerate,[11] the major legal demand for a forensic technique with demonstrably low conditional error probabilities with respect to the legal hypotheses is that the conclusion be reported with a suitable description of its probative value with regard to those hypotheses. In that way, scientists can inform the judge or judge without putting a heavy thumb on the scale for one side or the other. Within the match/no-match framework, this means that the ASTM standards must be expanded to address what is often called the "second stage" (Evett, 1990, 146 & 148; Garvin and Koons, 2011, 491; Parker, 1966, 38) - namely, determining the bearing of a match on the legal hypothesis that the questioned glass fragments came from the crime scene.[12]

At this second stage, there are two possibilities. If a matching rule with a well-defined significance level yields a nonmatch, life is (relatively) simple. The criminalist can report that the questioned specimens do not match according a rule that would catch at least the specified percentage of true matches. Inversely, if they do match, then the laboratory must estimate the probability of a match for questioned specimens originating from glass elsewhere in the population.[13] Thus, the fact that it is harder for a laboratory whose measurements are imprecise to reject H_0 is balanced by the fact that the laboratory cannot report as

[11] It has been argued that in civil cases, a rejection region should be defined so as to equalize Type I and Type II error probabilities. This notion does not implement the more-probable-than-not burden of persuasion for civil cases (Kaye, 1987), and it should not be confused with equivalence testing of drugs (Walker and Nowacki, 2011), which is a much more sensible approach to overcoming the inertia of the null hypothesis.

[12] The logic (and limitations) of the two-stage procedure are well understood in the forensic-science literature (e.g., Curran et al., 2000). Campbell and Curran (2009, 2) argue that the "two step approach is neither necessary nor desirable" although it is "the forensic norm".

[13] This is a Type II error probability for the legal same-source hypothesis. It is not a Type II error probability for the statistical hypothesis of "same RI" or "same elemental composition" that the significance level of the matching rule addresses.

impressive an 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-stage statistic that indicates the probative force of the matching measurements (Kaye, 1995).

This two-stage procedure is hardly ideal. It discards information (ibid.), and it risks confusion on the part of lawyers, judges, jurors, and even expert witnesses who must present or explain these probabilities (Kaye et al., 2011). But it is not legally foreclosed. In fact, it is the norm in the United States for reporting DNA results in most single-source cases, and it was the norm even when the variable being measured was essentially continuous. Gel electrophoresis of Variable Number Tandem Repeat (VNTR) alleles entailed length measurements that were unable to resolve the exact number of tandem repeats. The response of DNA testers was "matching" and "binning." The matching phase used wide windows for declaring a match the DNA fragments were "indistinguishable" on the basis of their positions on the gels according to the matching rule. A US National Academy of Sciences report (National Research Council, 1996) concluded that the two-stage methods for matching and binning that were in use were adequate because two things were known. First, the match criteria rarely failed to include truly matching fragments, and, second, the random-match probabilities computed with the bin frequencies could be presented to jurors to enable them to assess the significance of a finding that the fragments were "indistinguishable." The courts generally accepted matches determined in this fashion (Kaye, 2010).

Analogously, if the glass standards are to standardize the work of criminalists who use this mode of statistical inference, they must include advice on procedures for estimating and reporting how improbable false matches really are. Given research that already has been published or is underway, that may be possible, but this second stage of research has yet to be translated into standards that would inform and guide the practice of investigating the origin of glass fragments. It is time to move at least this far in the development of statistically thoughtful and well-founded forensic-science standards.

ACKNOWLEDGMENTS

I am grateful to José Almirall, Cédric Neumann, and Clifford Spiegelman for comments on a preliminary draft of this article. Naturally, the views expressed here are mine and do not necessary reflect the positions or opinions of these or any other individuals or organizations.

REFERENCES

Aitken, C.G.G., Roberts, P. and Jackson G. (2010). *Fundamentals of Probability and Statistical Evidence in Criminal Proceedings: Guidance for Judges, Lawyers, Forensic Scientists and Expert Witnesses*. Royal Statistical Society, London.

Aitken, C.G.G. and Taroni, F. (2004). *Statistics and the Evaluation of Evidence for Forensic Scientists*. 2d ed. John Wiley & Sons, Chichester.

Almirall, J. (2013). *Significance of Elemental Analysis From Trace Evidence: Final Technical Report For Grant Number 2009-DN-BX-K252 from the Department of Justice.* Office of Justice Programs, National Institute of Justice, Document No. 242325. National Criminal Justice Reference Service: Washington DC.

Almirall, J., and Trejos, T. (2016a). Analysis of glass evidence. In J.A. Siegel, editor, *Forensic Chemistry: Fundamentals and Applications.* John Wiley & Sons, Chichester: 228 – 27.

Almirall, J. (2016). Personal communication.

ASTM (2011). *E1967–11a Standard Test Method for the Automated Determination of Refractive Index of Glass Samples Using the Oil Immersion Method and a Phase Contrast Microscope*. ASTM Inc., West Conshohocken, PA.

ASTM (2012). E2330–12 *Standard Test Method for Determination of Concentrations of Elements in Glass Samples Using Inductively Coupled Plasma Mass Spectrometry (ICP-MS) for Forensic Comparisons*. ASTM Inc., West Conshohocken, PA.

ASTM (2013). *E2926–13 Standard Test Method for Forensic Comparison of Glass Using Micro X-ray Fluorescence (μ-XRF) Spectrometry*. ASTM Inc., West Conshohocken, PA.

Balding. D.H. and Steele, C. (2015). *Weight-of-Evidence for Forensic DNA Profiles*, 2d ed. John Wiley & Sons, Chichester.

Bottrell, M.C. (2009). *Forensic Glass Comparison: Background Information Used In Data Interpretation, Forensic Science Communications, 11.* https://www.fbi.gov/about-us/lab/forensic-science-communications/fsc/april2009/review/2009_04_review01.htm/.

Broun, K.S., Dix, G.E., Imwinkelreid, E.J., Kaye, D.H., Mosteller, R.P., Roberts, E.F. and Swift, E. (2013). *McCormick on Evidence*, vol. 1, 7th ed. Thompson Reuters, Eagan MN.

Buckleton, J.S., Bright, J. and Taylor, D. (2016). *Forensic DNA Evidence Interpretation*, 2nd ed. CRC Press, Boca Raton FL.

Carracedo, A. (2015). Forensic Genetics: History. In M.M. Houck, editor, *Forensic Biology*. Academic Press, Oxford, 19-22.

Campbell, G.P. and Curran, J.M. (2009). The interpretation of elemental composition measurements from forensic glass evidence III. In *Science and Justice* . 49: 2–7.

Curran, J.M., Hicks, N.T. and Buckleton, J.S. (2000). *Forensic Interpretation of Glass Evidence*. CRC Press, Boca Raton FL.

Dorn, H., Ruddell, D.E., Heydon, A. and Burton, B.D. (2015). Discrimination of float glass by LA-ICP-MS: Assessment of exclusion criteria using casework samples. In *Canadian Society Forensic Science Journal*. 48: 85–96.

Evett, I.W. (1990). The theory of interpreting scientific transfer evidence. In A. Maehly and R.L. Williams, editors, *Forensic Science Progress*. Springer Verlag, Berlin: vol. 4, 143–180.

Evett, I.W. and Weir, B.S. (1998). *Interpreting DNA Evidence: Statistical Genetics for Forensic Scientists*. Sinauer Associates: Sunderland MA.

Foster + Freeman Ltd. (2016). ffTA GRIM¨3 Glass RI Measurement, http://www.fosterfreeman.com/trace-evidence/304-ffta-grim.html.

Garvin, E. J. and Koons, R.D. (2011). Evaluation of match criteria used for the comparison of refractive index of glass fragments. In *Journal of Forensic Sciences*. 56: 491–500.

Gaudette, B.D. (1988). The forensic aspects of forensic textile examination. In R. Saferstein, editor, *Forensic Science Handbook*. Prentice-Hall, Englewood Cliffs NJ: vol. 2, 209–272.

Joint Committee for Guides in Metrology (2008). *International vocabulary of metrology—basic and general concepts and associated terms (VIM)*, 3d ed. Bureau International des Poids et Mesures: Paris.

Kaye, D.H. (1987). Hypothesis testing in the courtroom. In A. Gelfand editor, *Contributions to the Theory and Application of Statistics*, Academic Press, 1987: 331–356.

Kaye, D.H. (1987). Apples and oranges: Confidence coefficients and the burden of persuasion. In *Cornell Law Review*. 73: 4–77.

Kaye, D.H. (1995). The relevance of "matching" DNA: Is the window half open or half shut? In *Journal of Criminal Law and Criminology*. 85: 676–695.

Kaye, D.H. (2005). The NRC bullet-lead report: should science committees make legal findings? In *Jurimetrics Journal*. 46: 91–105.

Kaye, D. H. (2016). The interpretation of DNA evidence: a case study in probabilities, an educational module prepared for the Committee on Science, Technology, and Law, National Research Council.

Kaye, D.H. (2010). *The Double Helix and the Law of Evidence*. Harvard University Press, Cambridge MA.

Kaye, D.H. and Freedman, D.A. (2011). Reference guide on statistics. In *Reference Manual on Scientific Evidence*, 3d ed. National Academy Press, Washington DC: 211–302.

Kaye, D.H., Bernstein, D.E. and Mnookin, J.L. (2011). *The New Wigmore: A Treatise on Evidence: Expert Evidence*. New York: Aspen Publishing Co., New York.

Koons, R.D. and Buscaglia, J. (2002). Interpretation of glass composition measurements: the effects of match criteria on discrimination capability. In *Journal of Forensic Sciences*. 47: 505–512

Melsa, J.L. and Cohn, D.L. (1978). *Decision and Estimation Theory*. McGraw Hill: New York.

Microtrace (2016). *Glass Refractive Index Measurement System.* https://www.microtracellc.com/technique/glass-refractive-index-measurement-system-grim/.

Miller, E.T. (1982). Forensic glass comparisons. In R. Saferstein, editor, *Forensic Science Handbook*. Prentice-Hall, Englewood Cliffs NJ: vol. 1, 139–183.

National Research Council, Committee on the Evaluation of Forensic DNA Evidence: An Update (1996). *The Evaluation of Forensic DNA Evidence*. National Academy Press: Washington DC.

Parker, J.B. (1966) A statistical treatment of identification problems. In *Journal of the Forensic Science Society*. 6: 33–39.

President's Council of Advisors on Science and Technology (2016). Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature Comparison Methods. Executive Office of the President: Washington DC. www.whitehouse.gov/ostp/pcast

SWGMAT (Scientific Working Group for Materials Analysis) (2004). *Glass Refractive Index Determination*, available at https://www.fbi.gov/about-us/lab/forensic-science-communications/fsc/jan2005/index.htm/standards/2005standards9.htm.

Van Dyk, D. A. (2014). The role of statistics in the discovery of a Higgs boson. In *Annual Review of Statistics and Its Applications.* 1: 41–59.

Volokh, A. (1997). n guilty men. University of Pennsylvania Law Review 146: 173–216.

Walker, E. and Nowacki, A.S. (2011). Understanding equivalence and noninferiority testing. In *Journal of General Internal Medicine*. 26: 192–196.

Weis, P., Dücking, M., Watzke, P., Menges, S. and Becker, S. (2011). Establishing a match criterion in forensic comparison analysis of float glass using laser ablation inductively coupled plasma mass spectrometry. In *Journal of Analytical Atomic Spectrometry.* 26: 1273–1284.

Zadora, G., Martyna, A., Ramos, D. and Aitken, C. (2014). *Statistical Analysis in Forensic Science: Evidential Value of Multivariate Physicochemical Data*. John Wiley & Sons: Chichester.