Final Report

Agency: National Institute of Justice

Award No: 2015-DN-BX-K049

Title: Strengthening the evaluation and interpretation of glass evidence using statistical analysis of collection sets and databases of refractive index and elemental data (uXRF, ICP-MS and LA-ICP-MS)

Principal Investigator:
Dr. Jose Almirall, Professor
Professor, Department of Chemistry and Biochemistry and
Director Emeritus, International Forensic Research Institute (IFRI)
Director, Center for Advanced Research in Forensic Science (CARFS)
Florida International University
11200 SW 8th Street, OE116
Miami, FL 33199
(305) 348-3917 tel
(305) 348-4485 fax
almirall@fiu.edu

[redacted] [redacted]

[redacted] [redacted]

Recipient Organization Address:
Florida International University
11200 SW 8th Street
Miami, Florida, 33199

Submitting Official:
Mr. Roberto Gutierrez
Assistant Vice President for Research
Office of Research and Economic Development
11200 SW 8th Street, MARC 430
Miami, Florida, 33199
305-348-2494
gutierrr@fiu.edu

Project Period: 1/1/2016 to 12/31/2018

Report term: Final Report

Submission date: 05/30/2019

Submitting Official Signature:

Strengthening the evaluation and interpretation of glass evidence using statistical analysis of collection sets and databases of elemental data using LA-ICP-MS

Final Report by: Jose Almirall, Tricia Hoffman and Ruthmara Corzo
Department of Chemistry and Biochemistry, Florida International University, Miami, Florida – almirall@fiu.edu

Abstract

According to the National Highway Traffic Safety Administration Fatal Analysis Report System (NHTSA, 2018), ~6% of the 34,247 fatal crashes in the United States during 2017 involved a hit-and-run driver. These accidents resulted in ~2000 fatalities, a number that has seen a significant increase over the last decade. When drivers leave the scene of a hit-and-run collision without rendering aid and/or providing information to the others involved in the crash, the result is a crime scene with a variety of forensic evidence that can be used to reveal the parties involved. Trace evidence found at the scene of a violent collision between vehicles, or between a vehicle and a pedestrian, often provides leads that can assist in an investigation. While the victim may shed significant biological material, the nature of a hit-and-run investigation is that the vehicle must be located first, making timely leads from trace evidence pivotal. Plastic pieces of vehicle parts, paint chips and smears, glass shards, garment impressions, air bag residues, and other "trace evidence" is often more useful than biological evidence at the early stages of an investigation. Trace evidence may provide answers to pertinent questions such as what type of vehicle was involved, and ***who was driving*** at the time of the crash.

This research aims to improve on the value of glass evidence analysis by developing objective and quantitative interpretation guidelines for the evaluation and reporting of glass evidence analysis results. The first part of this research focused on the development of a reporting scheme based on previously reported likelihood ratio calculations by Aitken and Lucy [**1**] in conjunction with a "calibration" step previously reported by Ramos [**2**]. The aim of the first part of this research was to develop a database resulting from the analysis of glass using laser ablation inductively coupled plasma mass spectrometry (LA-ICP-MS). An important consideration in forensic chemistry is the interpretation of the evidence. Typically, a match criterion is used to compare the known and questioned sample. However, match criteria suffer from several disadvantages that can be overcome with an alternative approach: the likelihood ratio (LR). Two LA-ICP-MS glass databases were used to evaluate the performance of the LR: a vehicle windshield database (420 samples) and a casework database (385 samples). Compared to the match criterion, the likelihood ratio led to improved false exclusion rates (< 1.5%) and similar false inclusion rates (< 1.0%). In addition, the LR limited the magnitude of the misleading evidence, providing only weak support for the incorrect proposition [**3**]. The likelihood ratio was also tested through a series of three inter-laboratory studies including up to ten LA-ICP-MS forensic laboratory participants. Good correct association rates (94-100%) were obtained for same-source samples for all three inter-laboratory exercises. Moreover, the LR showed a strong support for an association. All different-source samples were correctly excluded with the LR, resulting in no false inclusions [**4**].

Introduction

An important consideration in trace analysis is the evaluation of the evidence. Usually, a match criterion (e.g., t-test, range overlap, *n*-sigma) is used to compare the known and questioned sample [**5-7**]. The two samples are considered to be indistinguishable if no differences in their elemental profile are found. On the hand, the known and questioned samples are considered to be distinguishable if at least one element is found to differ. The match criterion approach, referred to as the "frequentist approach," has several disadvantages: it suffers from the "fall off the cliff" effect, in which a small change in the evidence can lead to a drastic change in the final decision; it does not account for the rarity of an elemental profile; and it does not provide a weight of evidence [**8-11**]. The Bayesian approach is an alternative method for evidence interpretation that does not suffer from the disadvantages stated above. Bayes theorem is defined as:

$$\frac{P(H_1|E)}{P(H_2|E)} = \frac{P(E|H_1)}{P(E|H_2)} \times \frac{P(H_1)}{P(H_2)}$$

Hypothesis one, H_1, supports an association (i.e., no difference between the known and unknown sample), and hypothesis two, H_2, supports no association. The first term is known as the posterior odds, the middle term is the likelihood ratio (LR) and the right term is known as the prior odds. To a forensic scientist, only the likelihood ratio is typically of interest [**12**]. The LR is the ratio of the probability of the evidence (E) given H_1 divided by the probability of the evidence given H_2. A LR greater than 1 therefore supports H_1, while a LR less than 1 supports H_2; if the LR equals 1, neither hypothesis is supported. Moreover, a larger LR shows stronger support for an association and a smaller LR shows stronger support for an exclusion (non-association). Unlike the frequentist approach, the LR provides a quantitative and more objective approach to evidence interpretation. The LR has been applied to many types of forensic evidence including, but not limited to: glass, paint, gunshot residue, fingerprints, illicit drugs, DNA, and speaker recognition [**13-36**].

A total of 420 windshield glass samples were collected and analyzed using LA-ICP-MS. Additionally, a 385-sample casework glass database was provided by the Bundeskriminalamt (BKA) in Germany. The quantitative data for each database was used to calculate a likelihood ratio through a cross validation study. This research presents the first study to directly compare the match criterion currently in use in the United States to the relatively new likelihood ratio. Additionally, an inter-laboratory study was conducted in order to test the performance of the likelihood ratio in mock case scenarios. The inter-laboratory study was part of an ongoing effort to standardize the interpretation of forensic evidence as well as the reporting language used in case reports.

The Natural Isotopes and Trace Elements in Criminalistics and Environmental Forensics (NITECRIME) European Network developed a quantitative methodology for the analysis of glass fragments using LA-ICP-MS [**37,38**]. The group also developed two new float glass standards (FGS 1 and FGS 2) that served as matrix-matched standards for the analysis of soda-lime glass using LA-ICP-MS. Trejos, et al. conducted a comprehensive study comparing µXRF, solution ICP-MS, laser ablation (LA) ICP-MS, and LA-ICP-OES as part of the Elemental Analysis Working Group (EAWG) [**5,39**]. Both µXRF and ICP-based methods performed well in terms of accuracy and precision using glass standards (NIST 612, NIST 1831, FGS 1, and FGS 2). Moreover, all participating laboratories correctly associated same-source and correctly discriminated different-source glass samples that were submitted as mock casework to each lab. As expected ICP-based techniques provided superior sensitivity; the µXRF limits of detection (LODs) were typically 2 to 3 orders of magnitude greater than those of ICP techniques. Still, the authors concluded that both µXRF and ICP-based techniques are fit-for-purpose for the forensic analysis of glass. The evaluation of the performance for several match criteria ultimately led to a standard methodology for µXRF and LA-ICP-MS (ASTM E2926 and ASTM E2927, respectively) [**40-41**].

The performance of match criteria using elemental data in order to distinguish different-source or associate same-source glass fragments has been extensively researched [**5-7, 40-41, 42**]. Koons and Buscaglia analyzed 209 glass fragments using Inductively Coupled Plasma-Atomic Emission Spectroscopy (ICP-AES) [**7**]. The equal-variance T-test (at 95% confidence and with the Bonferroni correction) led to 2 false inclusions, resulting in a low false inclusion rate of 0.009%. When using the unequal variance T-test, as opposed to equal variance, a higher false inclusion rate of 0.055% was obtained (12 falsely included pairs). On the other hand, using range overlap resulted in no false inclusions. The authors concluded that either tool (i.e., Bonferroni-corrected T-test or range overlap) is appropriate for the comparison of glass fragments. However, it should be noted that same-source comparisons were not included in this study; thus, the false exclusion rate of each statistical tool is not investigated.

This resource was prepared by the author(s) using Federal funds provided by the U.S.
Department of Justice. Opinions or points of view expressed are those of the author(s) and do not
necessarily reflect the official position or policies of the U.S. Department of Justice.

A larger study, the Elemental Analysis Working Group (EAWG), investigated the performance of several match criteria using elemental data collected using X-ray Fluorescence Spectroscopy (XRF) and ICP techniques (ICP-MS, LA-ICP-MS, ICP-AES) [**39**]. Mock casework samples were sent to each of the 9 XRF labs and 7 ICP labs as part of an inter-laboratory study. Each lab was asked to analyze the glass fragments they received and compare the data using the following match criteria: range overlap, T-test (99% confidence and 95% confidence with and without the Bonferroni correction), Hotellings T^2, and ± 2, 3, 4, 5, and 6 standard deviations (SD). For pairwise comparisons using XRF data, range overlap and ±3 SD offered the best compromise between the false exclusion and false inclusion rates: :: 19% and :: 27% respectively. For ICP methods, a modified ± 4 SD, using a minimum SD equal to 3% of the average, performed best (:: 28% false exclusion rate and :: 5% false inclusion rate). The results of the EAWG study ultimately led to two standard methodologies: one for the analysis of glass using XRF (ASTM E2926) and the other for the analysis of glass using LA-ICP-MS (ASTM E2927) [**40,41**].

A separate European Working Group, the Natural Isotopes and Trace Elements in Criminalistics and Environmental Forensics (NITECRIME), developed a quantitative methodology for the analysis of glass fragments using solution LA-ICP-MS [**37,38**]. Using the quantitative method developed, Weis et al. investigated a match criterion that takes inter-day variation into account. Two datasets were used to calculate the false exclusion and false inclusion rate, respectively: a single glass pane analyzed 44 times (6 replicates each) and a set of 62 different-source float glass samples. A control sample (DGG 1) was analyzed 90 times and the overall relative standard deviation (RSD) was calculated for each element, which was then used as a "fixed relative standard deviation" (FRSD) for pairwise comparisons. The comparison interval for the known sample was defined by an upper limit of the known average × (1 + *n* × FRSD) and a lower limit of the known average ÷ (1 + *n* × FRSD). The *n* indicates the sigma value used; sigma values of 1-6, 8, 10, 15, and 20 were tested to determine the best compromise between the false exclusion (Type I) and false inclusion (Type II) rate. The 4-sigma match criterion performed best, with a Type I error of 14.83% and no false inclusions. Although casework samples are typically analyzed on the same day, accounting for inter-day variation is beneficial in order to establish a random match probability or a frequency. In the former case, the false inclusion rate for all possible comparisons using a glass database is reported; the random match probability gives an indication of the probability of coincidental "matches" between glass fragments of different origin. In the latter case, the questioned sample from casework is compared to all samples in a database and the number of "matches" is reported; the frequency provides an estimation of the rarity of a particular elemental profile (i.e., that of the questioned sample).

Several criticisms regarding the frequentist approach (i.e., match criterion) have been reported. First, it suffers from the "fall off the cliff" effect, in which a small change in the significance value, *p*, leads to a drastic change in the interpretation of the data. The "fall off the cliff" effect occurs when the average of the questioned sample lies close to the cut-off established by the known sample's comparison interval. For example, if the known comparison interval for one element was 5 – 10 parts per million (ppm) and the average of the questioned sample was 9.99 ppm, the K and Q would be considered indistinguishable; but if the Q average was slightly higher (e.g., 10.01 ppm), the K and Q would be distinguishable. A second disadvantage of the frequentist approach is that it does not take the rarity of the elemental profile into account. This second drawback may be overcome by using a glass database to calculate a frequency; however, the use of a database to generate a frequency does not eliminate the "fall off the cliff" effect. Finally, the frequentist approach answers the "pre-data" rather than the "post-data" question. The former answers the question "what is the probability of a match if I carry out this procedure." The latter answers the question that the court is interested in: "how much does this evidence increase the likelihood that the suspect is guilty." [**1, 8-10**]

There are numerous methods for calculating the likelihood ratio. The simplest approach is to calculate a frequency, *f* = *n* ÷ *N*; *n* is the number of times the questioned sample "matches" a sample in the database (including the known sample) and *N* is the total number of samples in the database. The likelihood ratio (LR) is often estimated as the reciprocal of the frequency: LR = 1 ÷ *f* = *N* ÷ *n*. However, the denominator of the LR should evaluate the number of alternative sources and should therefore be estimated as: LR = *N* ÷ (*n* – 1) [**43**]. The frequency approach is only used when the known (K) and questioned (Q) sample are found to be indistinguishable; if the two samples are distinguishable, then the K and Q are excluded and no LR is calculated. An advantage of this approach is its simplicity. Nevertheless, since the frequency approach uses a match criterion for pairwise comparisons, it still suffers from the "fall off the cliff" effect.

Aitken and Lucy proposed a Multivariate Kernel (MVK) model [**1**] that accounts for two levels of variation in multivariate data: the within-source variation (multivariate normal) and the between-source variation (KDE). An alternative approach, the Multivariate Normal model (MVN), assumes multivariate normality for both the within-source and between-source variation. The authors recommend the MVK model but state that if the between-source distribution is well represented by a multivariate normal distribution, then the MVN model may perform as well as the MVK model. The LR calculation using either the MVK or MVN models can be implemented using the freely available R packages: "comparison," "nnls," and "isotone."[**44-46**]

Unfortunately, the MVK model can lead to extremely large or small LRs, providing an unreasonable weight of evidence; this is the case for LA-ICP-MS glass data. The extreme LRs are likely a result of the high dimensionality of the data (i.e., many variables). Thus, a post-hoc calibration may be necessary in order to reduce the feature-based LR to more reasonable values. Calibration is accomplished by treating the LR as a score, rather than an actual likelihood ratio, and then transforming the score into a LR.

Vergeer, et al. and van Es, et al. reported one method for calibration that involves the use of density models followed by the empirical lower and upper bound (ELUB) method to limit the LR output [**13-14**]. The distribution of the same-source LR *scores* (using the MVK model) was modeled using a double exponential decay and the distribution of the different-source LR *scores* was modeled using a KDE [**13**]. To compute the calibrated LR for a pairwise comparison, first the LR score is calculated using the MVK model. The numerator of the calibrated LR is given by the probability of the score using the same-source distribution (in this case, a double exponential decay). The denominator of the calibrated LR is given by the probability of the score using the different-source KDE. The upper and lower limit for the calibrated LR is computed using a normalized Bayes error-rate (NBE) plot, which plots the $\log_{10}$ EU ratio against the $\log_{10} LR_{th}$.[**14**]

Another method for LR calibration employs the Pool Adjacent Violators (PAV) algorithm, which uses strictly proper scoring rules (SPSRs) [**11, 47-49**]. The ELUB method described above includes one step for calibration and a subsequent step to limit the LR. The PAV transformation, on the other hand, simultaneously calibrates and sets an upper and lower limit to the LR, while still maintaining the discriminating power of the LRs [**50,51**]. The algorithm gives a non-decreasing transformation for each posterior probability (corresponding to each un-calibrated LR) such that the resulting posterior probabilities are better calibrated. Recall that Bayes theorem is defined as: posterior odds = likelihood ratio × prior odds.

$$\frac{P(H_1|E)}{P(H_2|E)} = \frac{P(E|H_1)}{P(E|H_2)} \times \frac{P(H_1)}{P(H_2)}$$

A FIU vehicle glass collection was created as a result of this research and contains a total of 420 glass samples taken from the inner and outer windshield pane of 210 vehicles located at the M&M Service and Salvage Yard in Ruckersville, Virginia. A complete list of all vehicles sampled is provided in Corzo et al. [**3**]. Windshield glass consists of two glass panes held together by a plastic film. The windshield glass was cut using a RHYNO laminated glass cutter. One to three large (~6 by 8 cm) glass pieces were collected depending on how much glass was available for a particular vehicle. Three small pieces (typically less than 1 cm^2) were taken from the outer pane and the side of the pane that was not in contact with the polymer film and the same was done for the inner pane. In some cases, this pane side corresponded with the float side (verified by the large ^{118}Sn

signal). For each of the three small fragments, 5 replicate measurements were collected, for a total of 15 replicates per sample.

The 17 isotopes listed in ASTM E2927 were monitored: ^{7}Li, ^{25}Mg, ^{27}Al, ^{39}K, ^{42}Ca, ^{49}Ti, ^{55}Mn, ^{57}Fe, ^{85}Rb, ^{88}Sr, ^{90}Zr, ^{137}Ba, ^{139}La, ^{140}Ce, ^{146}Nd, ^{178}Hf, and Pb (average of ^{206}Pb, ^{207}Pb, and ^{208}Pb).^{115} Two additional isotopes were monitored, but not used for characterization purposes: ^{29}Si was used as the internal standard since SiO_2 is present at high concentration (~ 72%) and ^{118}Sn was monitored to determine the float side of the glass pane. The isotopes were quantified using single-point calibration with the Glitter™ software (MacQuarie University, Australia). Float Glass Standard 2 (FGS 2) was used as the calibrator, while FGS 1 and NIST 1831 were analyzed daily to assess bias. Prior to analysis, a daily performance using the NIST 612 standard was performed to ensure that the instrument sensitivity was adequate and that doubly charged species as well as oxides were below 3%. A ns-213 nm Nd:YAG laser (ESI New Wave Research, Portland OR USA) coupled to a quadrupole ELAN DRC II (Perkin Elmer LAS, Shelton CT USA) was used for analysis. The laser parameters for analysis were as follows: 100% laser energy (~ 0.65 mJ), 10 Hz, 90 µm spot size, and 60-second dwell. The laser ablation parameters and the ICP-MS method were previously developed during two inter-laboratory studies in which FIU participated: the Natural Isotopes and Trace Elements in Criminalistics and Environmental Forensic (NITECRIME) and the Elemental Analysis Working group (EAWG) [**5, 37, 39, 41, 42**]. In general, 8 vehicles were completed in one day (approximately a 12 hour run). This amounts to 48 glass fragments in one day (excluding the calibrator and controls): 3 fragments for each inner and outer pane of every vehicle. The calibrator (FGS 2) was analyzed at the beginning, middle, and end of the sequence in order to account for instrumental drift. NIST 1831 was analyzed in the first half of the sequence, while FGS 1 was analyzed in the second half of the sequence. After analysis of the 420 glass samples (inner and outer panes from 210 vehicles), 40 randomly selected duplicates were reanalyzed in order to assess the correct association rate. All glass samples and selected duplicates were analyzed over a total of 40 days. In order to assess the spread of each variable (element) in the FIU vehicle database, a box and whisker plot was produced (Figure 1, left). In the plot, the $\log_{10}$ of the concentration was used so that all elements could be plotted in scale.

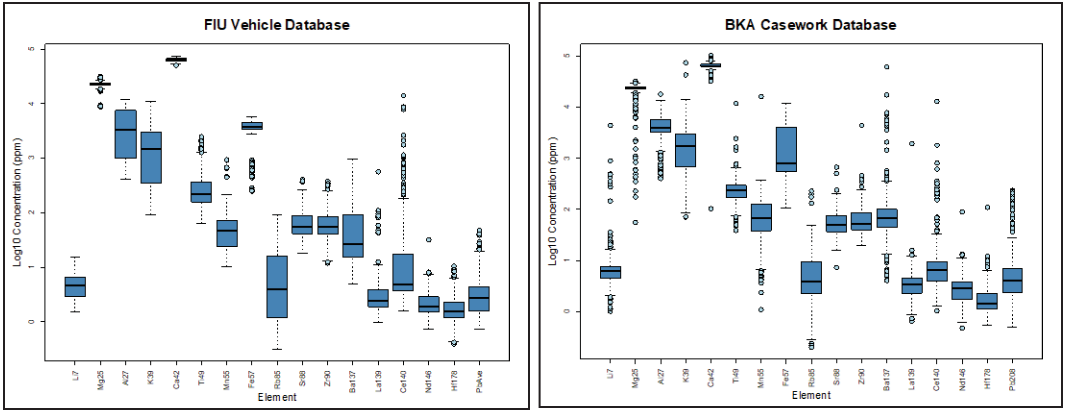

Figure 1. Box and whisker plots for FIU vehicle database (left) and the BKA database (right).

The horizontal black line across each box represents the median value, while the upper and lower limits of the box are the upper and lower quartiles (75th and 25th percentile), respectively. The upper and lower whiskers are calculated as follows:

$$\mathit{Upper\ whisker} = Q1 + 1.5 \times \mathit{!QR\ and\ Lower\ whisker} = Q1 - 1.5 \times \mathit{!QR}$$

In the equations above, *Q3* is the upper quartile (i.e., the upper limit of the box), *Q1* is the lower quartile (i.e., the lower limit of the box), and *IQR* is the box length (*Q3 – Q1)*. If the calculation of the upper whisker, using the formula above, is greater than the maximum value in the database, then the upper whisker is set as the maximum value; likewise, if the calculation of the lower whisker is less than the minimum value in the database, then the lower whisker is set as the minimum value. The light blue points that fall outside the whisker limits are extreme values that extend beyond the majority of the data for each particular element.

Some elements (e.g., Mg and Ca) show little variation across all database samples, as indicated by the small range of the box. Other elements (Al, K, Rb) have much wider spreads. Fe has many extreme values that are approximately an order of magnitude lower than the rest of the data. Ce has the opposite trend: the majority of samples have a low Ce concentration, but many extreme values have a concentration that is two or three orders of magnitude higher; a similar trend is seen for La. The remaining elements have a relatively moderate spread with few, or no, extreme values. In addition to the spread of each variable, the correlation between variables was investigated. Hf and Zr are nearly perfectly correlated (positive correlation of 0.996), thus it was not necessary to include both for the purpose of discrimination. K and Al and Ba and Pb are highly correlated as well (each pair has a positive correlation greater than 0.8). Other element pairs with fairly high correlation (positive correlation greater than 0.7) include: Al and Rb, K and Ba, and La and Nd.

The BKA Casework Database is comprised of 385 glass samples that were submitted as part of a case (both known and questioned samples). These samples were analyzed over a long period of time: from 2005 to 2016. The casework database includes an assortment of glass types: float glass, container glass, pre-float window glass, etc. As such, the assumption that SiO_2 is present at ~72% may not be valid. A box and whisker plot was also produced for the BKA casework database (Figure 1, right). Compared to the FIU database, the BKA database exhibits more extreme values. This is not surprising since the BKA database consists of many types of glass samples, while the FIU database only includes float glass taken from automobile windshields. Furthermore, the FIU database contains many samples from vehicles of the same make, model, and year of manufacture; samples from similar vehicles are expected to have similar elemental profiles. For the BKA casework database, Zr and Hf were found to be perfectly correlated (correlation of 1.0). The ASTM comparison criterion, as described above, was applied to all pairwise comparisons for the FIU vehicle database. Since this comparison criterion is asymmetrical, different results may be obtained depending on which sample is treated as the known or questioned sample. Thus, each pair of samples was compared twice using the comparison criterion. For example, Sample 1 (as the known) was compared to Sample 2 (as the questioned) using all 15 replicates of each sample; then the roles were reversed. This brings the total number of pairwise comparisons to:

$$n \times (n - 1)$$

In the equation above, n is the number of samples in the database. For the FIU database, n = 420 and the number of pairwise comparisons is therefore equal to 175,980. However, in some cases, the inner and outer pane of a vehicle windshield is expected to originate from the same manufacturing source. To account for this, any comparison between the inner and outer pane of the *same* vehicle was removed, bringing the total number of comparisons down to 175,560 (175,980 – 420). This ensures that only comparisons between *different* vehicles are treated as different sources.

Table 1 shows the false inclusion rate for the FIU database. Only 208 pairs out of the 175,560 totals pairs were associated, giving a low false inclusion rate of 0.12%. Of those 208 false inclusions, 165 were comparisons between vehicles of the same make and/or year of manufacture. This may explain why these pairs have similar elemental profiles and were found to be indistinguishable using the ASTM comparison criterion. Unfortunately, without the windshield sticker, the glass manufacturing plant where the windshield glass was produced is unknown. Therefore, it cannot be determined whether these false inclusions are truly random matches or correct associations. Still, even if all false inclusions truly originated from different sources (i.e., different glass manufacturers), the false inclusion rate is quite low.

Table 1. - False inclusions/exclusions for FIU database /duplicates using the ASTM comparison criterion.

	FIU Vehicle Database ASTM Comparison Criterion
False Inclusions	**0.12%** (208/175,560)
False Exclusions	**52.5%** (42/80)

In order to assess the false exclusion rate, the 40 duplicates were compared to their respective original sample. That is, Sample 1 was compared to Sample 1 Duplicate, Sample 2 to Sample 2 Duplicate, and so on. As mentioned previously, to account for the asymmetry of the ASTM comparison criterion, each pair was compared twice so that each sample was treated as the known and questioned. Thus, the total number of comparisons is equal to 80 (40 × 2). The ASTM criterion led to an extremely high false exclusion rate of 52.5%. The most discriminating elements were Al, Zr, and Pb. Al was especially problematic for duplicates analyzed on Day 44, which is not surprising since it is suspected that Day 44 suffered from a pulse/analog calibration issue. Some of the discriminating elements (e.g., La, Nd, Hf, Pb) are present at low concentrations. Other elements may have had high inter-day variation that was not accounted for using the ASTM criterion. For example, the BKA found that Zr and Hf have high inter-day variation. In order to account for this, the BKA established a fixed relative standard deviation (FRSD) that differs for each element.[37] We implemented an approach similar to that of the BKA; the results showed an improved false exclusion rate (1.9%, compared to 52.5% for the ASTM criterion). High inter-day variation might not affect casework, since the known and questioned samples are typically analyzed on the same day, as recommended in ASTM E2927. However, it would affect the use of a database to calculate a frequency of occurrence. For this approach, the questioned sample is compared not only to the known but also to an entire glass database using the ASTM comparison criterion. Then, the number of "matches" is counted (i.e., the number of times the questioned sample is indistinguishable to a sample in the database). The frequency of occurrence is subsequently calculated by:

$$Frequency = \frac{M}{N}$$

M is the number of "matches" and *N* is the total number of samples in the database (including the known). With the ASTM criterion, it is expected that the frequency of occurrence will be underrepresented. To correct this, a match criterion that accounts for the inter-day variation should be used. For this reason, if a laboratory aims to use a frequency of occurrence approach, the BKA approach is suggested (after the lab establishes its own FRSD). The ASTM criterion was also used to determine the false exclusion rate for the BKA casework database. For the BKA database, *n* = 385 and the number of comparison pairs is equal to 147,840. A false exclusion rate of 0.018% (28/147,840 pairs) was obtained. The much lower false exclusion rate for the BKA database, compared to the FIU database, is unsurprising because the FIU database contains samples from similar sources (i.e., same vehicle make/model or year of manufacture), whereas the BKA database includes casework samples of many different glass types. The false inclusion rate could not be estimated for the BKA database since no duplicate samples were analyzed.

The frequentist approach has several disadvantages: it suffers from the "fall-off-the-cliff" effect, in which a small change in the evidence can lead to a drastic change in the final decision; it does not account for the rarity of an elemental profile; and it does not provide a weight of evidence. The latter disadvantage, however, may be overcome through the use of a verbal scale in order to assign the strength of an association. Nonetheless, the use of a verbal scale can be subjective since it relies on the analyst's personal experience. A more objective approach is possible if a database is available.

The MVK model is also referred to as the two-level model since it accounts for two levels of variation: the within-source and between-source variation. The model uses a normal distribution for the within-source variation and a kernel density estimate for the between-source variation. A detailed description of the calculation for the numerator and denominator of the likelihood ratio using this model is given in Aitken and Lucy [**1**]. Unfortunately, the model leads to an unreasonable weight of evidence for LA-ICP-MS glass data; that is, the model results in extremely large or small LRs. This is likely due to the high dimensionality of the data (i.e., many variables). One method of calibration employs the Pool Adjacent Violators (PAV) algorithm. The PAV transformation improves the calibration of a set of posterior probabilities and sets an upper and lower limit to the LR, while still maintaining the discriminating power of the LRs.[140, 142] A detailed description of the algorithm is given in [**2**]. The PAV transformation has previously been applied to speaker recognition data and SEM-EDX glass data [**47-49**]. However, this study presents the first use of the PAV algorithm to calibrate LRs generated for LA-ICP-MS glass data. This calibration approach will subsequently be referred to as the “PAV method.” The performance of the likelihood ratio can be evaluated through the use of Empirical Cross Entropy (ECE) plots, which are described in [**2 and 52**].

Prior to the likelihood ratio calculation, all elements were normalized to the element with the highest average concentration (calcium). Then the base-10 logarithm was taken for the 16 element ratios. This data pre-treatment reduces the dimensionality and makes the multivariate probability distribution more amenable in subsequent calculations. A double 10-fold cross validation using the MVK model and PAV calibration was implemented in Matlab by Javier Franco-Pedroso and Daniel Ramos using the databases collected at FIU and the BKA; the results of this study have been described in detail by Corzo et al. **[3]**. Using the BKA database as the background population to calculate LRs for the FIU database is preferable since the BKA database includes real-world samples from actual cases submitted to the laboratory. On the other hand, the FIU database is comprised of glass from vehicles manufactured within a narrow time frame, many of which have the same make and/or year of manufacture. The BKA database therefore provides a more relevant population than the limited FIU database. Additionally, using the same database to train both the MVK and PAV models via a double cross validation may lead to overly optimistic results. The cost log-likelihood ratio (Cllr), which is a measure of accuracy, is the ECE value where the red line crosses a log10 prior odds equal to zero (x = 0). The closer the Cllr is to zero, the better the accuracy. The minimum cost log-likelihood ratio ($Cllr_{min}$), which is a measure of discrimination, is the ECE value where the blue line crosses a log10 prior odds equal to zero. The closer the Cllr is to zero, the better the discrimination. Finally, the calibration cost log-likelihood ratio ($Cllr_{cal}$) is the difference between Cllr and $Cllr_{min}$; the smaller the difference, the better the calibration. These three metrics are useful for the relative comparison of different LR systems.

Table 2 – False exclusion rate, false inclusion rate, Cllr and $Cllr_{min}$ (if applicable) for the BKA and FIU databases using different comparison criteria. Fractions within parentheses indicate the number of pairs that were falsely excluded/included over the total number of pairwise comparisons.

	BKA Database		***FIU Database***		
	ASTM Criterion	**LR (10-fold)**	**ASTM Criterion**	**LR (10-fold)**	**LR (with BKA)**
% False Exclusions	**3.25%** (25/770)	**0.52%** (2/385)	**7.50%** (63/840)	**1.19%** (5/420)	**1.19%** (5/420)
% False Inclusions	**0.022%** (33/147840)	**0.21%** (15/7220)	**0.10%** (88/87780)	**0.60%** (25/4200)	**0.33%** (143/43890)
Cllr	-	0.014	-	0.056	0.067
$Cllr_{min}$					
$Cllr_{cal}$	-	0.009	-	0.008	0.049

Table 2 shows the error rates for the FIU and BKA database using both the calibrated likelihood ratio and the ASTM comparison criterion. In order to fairly compare the two approaches, the pairwise comparisons using the ASTM criterion were redone so that the first 3 replicates of one sample were compared to the last 3 replicates of the same sample (for same-source comparisons) or a different sample (for different-source comparisons). Additionally, for the FIU database, all comparisons between inner and outer panes were excluded. Note that the total number of pairwise comparisons for the ATSM criterion is double that of the likelihood ratio; this is because the likelihood ratio is symmetrical, while the ASTM criterion is not.

Compared to the calibrated likelihood ratio, the ASTM criterion led to a higher false exclusion rate: 3.25% for the BKA database and 7.50% for the FIU database. It should be noted, however, that ASTM E2927 suggests a minimum of 9 replicate measurements to fully characterize the known sample, but only 3 replicate measurements were used in this study. It is possible that the high false exclusion rate is due to insufficient replicate measurements to fully characterize the known sample. The calibrated likelihood ratio offered a lower false exclusion rate: 0.52% for the BKA database and 1.19% for the FIU database (regardless of the background database used).

On the other hand, the false inclusion rate was lower for the ASTM criterion than the calibrated likelihood ratio. For the BKA database, a total of 31 pairs were associated using the ASTM criterion, leading to a low false inclusion rate of 0.022%. Using the calibrated LR, 15 pairs were associated yielding a false inclusion rate that was an order of magnitude greater than that of the ASTM criterion. Nonetheless, in both cases, the false inclusion rate is quite low (< 0.5%). Unfortunately, since the samples in the BKA database are casework samples, their origin is unknown. Therefore, it cannot be stated whether the two samples being compared were manufactured in the same glass plant at around the same time or whether, despite originating from different sources, the two samples have a chemical profile that coincidentally "match."

For the FIU database, the ASTM criterion resulted in a low false inclusion rate of 0.10%. However, though higher, the calibrated LR still performed well (< 1%) regardless of which database was used as the background population.

Apart from the rates of misleading evidence (false exclusions and inclusions), other metrics to assess the performance of the likelihood ratio include the previously mentioned log-likelihood ratio cost (Cllr) and the minimum log-likelihood ratio cost (Cllr$_{min}$) [**52**]. Table 30 shows a Cllr of 0.014 and a Cllr$_{min}$ of 0.049 (both values close to zero) for the BKA database, indicating both good accuracy and discrimination. Slightly worse accuracy and discrimination is seen for the two FIU database approaches. However, the Cllr and the Cllr$_{min}$ are still fairly close to zero for both cases. The calibration cost log-likelihood ratio (Cllr$_{cal}$) can be calculated by subtracting the Cllr$_{min}$ from the Cllr. The Cllr$_{cal}$ values are 0.008, 0.009, and 0.05 for the double 10-fold FIU database, the double 10-fold BKA database, and the FIU database with the BKA database as the background, respectively. Thus, the double 10-fold BKA and FIU database are approximately equally well calibrated (though the BKA database has better accuracy and discrimination). The FIU database with BKA background is well calibrated, but not as well as the other two approaches.

Table 3 – Percent of falsely inclusions that originate from the same vehicle make and/or year of manufacture for the FIU database. The fractions within parentheses indicate the number of false inclusions over the total number of different-source pairwise comparisons.

	ASTM Criterion	LR (10-fold)	LR (with BKA)
Same Vehicle Make and Year	**0.79%** (57/7220)	**0.48%** (20/4200)	**0.12%** (53/43890)
Same Vehicle Make, Different Year	**0.014%** (1/7220)	**0.048%** (2/4200)	**0.027%** (12/43890)
Same Year, Different Vehicle Make	**0.22%** (16/7220)	**0.024%** (1/4200)	**0.071%** (31/43890)
Different Vehicle Make and Year	**0.19%** (14/7220)	**0.048%** (2/4200)	**0.12%** (47/43890)

Since all of the FIU glass samples originated from known vehicles, it is possible to determine whether the glass from two different vehicles may have been produced in the same glass manufacturing plant. Table 3 shows the percentage of falsely included pairs that originated from similar vehicles. The greatest percentage of falsely included pairs originated from vehicles that have the same make and year of manufacture, thus it is likely that the windshield glass for these vehicles were produced in the same glass manufacturing plant and are therefore indistinguishable. In some cases, different automobile manufacturers obtain their glass from the same glass manufacturer; this may account for the falsely included pairs that have a different make but same year of manufacture. Relatively few false inclusions had the same vehicle make but different year of manufacture; for the most part, these vehicles were manufactured 1 to 2 years apart. The remaining pairs had a different vehicle make and year of manufacture; these too were typically manufactured 1 to 2 years apart. The highest number of false inclusions that originated from a different vehicle make and different year of manufacture can be seen for the calibrated LR using the BKA database as the background population. These pairs may genuinely be coincidental "matches" or, despite having different vehicle manufacturers, may have the same glass manufacturer. Unfortunately, although the vehicle origin for each glass sample is known, the windshield sticker that discloses the glass manufacturing plant was absent or illegible for most vehicles. Without this sticker, it is unknown whether the same glass manufacturer produced the windshield glass for these falsely included pairs. Many of the falsely excluded pairs using the ASTM criterion were discriminated by Pb, Li, and/or Hf, all of which are typically present at low concentrations (< 3 ppm) and are therefore near the limit of detection. The remaining false exclusions were distinguished by: Zr, Nd, Sr, La, Ce, Ba, Rb, Mn, and/or Ti. This may be because the ASTM criterion does not account for large (> 3% relative standard deviation) inter-day variation. Rather than using a minimum of 3% of the average, it may be beneficial to establish a different minimum depending on the inter-day variation for each particular element (the approach that the BKA applies for their match criterion) [**6**]. Some of the false exclusions were due to the "fall-off-the-cliff" effect, in which the average concentration of the questioned sample fell just outside the comparison interval for the known sample. For example, one same-source pair was discriminated by Ba and Pb. For both elements a difference of less than 0.2 ppm for the average concentration of the questioned sample would have led to a correct association. The calibrated likelihood ratio correctly associated most of the falsely excluded pairs that were due to the "fall-off-the-cliff" for the ASTM criterion.

Overall, the calibrated likelihood ratio performed well for the double 10-fold cross validation experiments, for the smaller independent datasets, and for the inter-laboratory data. Moreover, it is clear that the LR provides several advantages over the match criterion: it provides a quantitative measure for the weight of evidence, it does not suffer from the "fall off the cliff" effect, and it takes the rarity of the elemental profile into account. Still, the results indicate that the selection of the background database is an important step for the calculation of the LR. A disadvantage of the LR approach is that many forensic laboratories do not have a database available. It may be possible for forensic glass analysts to use a single compiled database for the calculation of the LR. A shared database may not be the most ideal approach since the frequency of glass elemental profiles is expected to differ across different locations and thus may not be representative of the relevant population. However, this study showed that regardless of the database used, good correct association rates (> 94%) and no false inclusions were obtained, thus the use of a shared database may be justified. A second disadvantage of the LR approach is its complexity. An exit survey was sent to all inter-laboratory participants in order to gauge their reactions to the LR approach, which is a relatively new approach in the United States. Several participants stated that the following would hinder their use of the LR approach: the complexity of the calculations, the difficulty in interpreting the LR value, the difficulty in explaining the LR in court, and the opinion that the current method (i.e., the match criterion) is appropriate and the LR is unnecessary. Therefore, before the LR is widely accepted in the forensic community, the analysts must be convinced of the improvement the LR offers over the currently used, and more subjective, interpretation approaches (e.g., verbal scale, frequency). The development of a user-friendly program for the calculation of the LR that is accessible to practitioners can potentially aid in encouraging analysts to become more familiar with the LR approach. Moreover, a standard methodology for the calculation of the LR is expected to increase the acceptance of the LR approach within the U.S. forensic community.

LIST OF REFERENCES

1. Aitken, C. G. G.; Lucy, D., Evaluation of trace evidence in the form of multivariate data. *Journal of Applied Statistics* **2004,** *53*, 109-122.
2. Zadora, G.; Martyna, A.; Ramos, D.; Aitken, C., *Statistical Aalysis in Forensic Science: Evidential Value of Multivariate Physicochemical Data*. John Wiley & Sons, Ltd: West Sussex, United Kingdom, 2014.
3. Corzo, R.; Hoffman, T.; Weis, P.; Franco-Pedroso, J.; Ramos, D.; Almirall, J., The Use of LA-ICP-MS Databases to Estimate Likelihood Ratios for the Forensic Analysis of Glass Evidence. *Talanta* **2018,** https://doi.org/10.1016/j.talanta.2018.02.027.
4. T Hoffman, R Corzo, P Weis, E Pollock, A v Es, W Wiarda, A Stryjnike, H Dorne, A Heydon, E Hoise, S Le Franc, X Huifang, B Pena, T Scholz, J Gonzalez, JR Almirall, An Interlaboratory Evaluation of LA-ICP-MS Analysis of Glass and the Use of a Database for the Interpretation of Glass Evidence, *For. Chem.*, **2018** (11) 65-76. 10.1016/j.forc.2018.10.001
5. Trejos, T.; Koons, R.; Weis, P.; Becker, S.; Berman, T.; Dalpe, C.; Duecking, M.; Buscaglia, J.; Eckert-Lumsdon, T.; Ernst, T.; Hanlon, C.; Heydon, A.; Mooney, K.; Nelson, R.; Olsson, K.; Schenk, E.; Palenik, C.; Pollock, E. C.; Rudell, D.; Ryland, S.; Tarifa, A.; Valadez, M.; van Es, A.; Zdanowicz, V.; Almirall, J., Forensic analysis of glass by μ-XRF, SN-ICP-MS, LA-ICP-MS and LA-ICP-OES: evaluation of the performance of different criteria for comparing elemental composition. *Journal of Analytical Atomic Spectrometry* **2013,** *28* (8), 1270-1282.
6. Weis, P.; Dücking, M.; Watzke, P.; Menges, S.; Becker, S., Establishing a match criterion in forensic comparison analysis of float glass using laser ablation inductively coupled plasma mass spectrometry. *Journal of Analytical Atomic Spectrometry* **2011,** *26* (6), 1273-1284.
7. Koons, R. D.; Buscaglia, J., Interpretation of Glass Composition Measurements: The Effects of Match Criteria on Discrimination Capability. *Journal of forensic sciences* **2002,** *47* (3), 505-512.
8. Curran, J. M.; Triggs, C. M.; Almirall, J. R.; Buckleton, J. S.; Walsh, K. A. J., The interpretation of elemental composition measurements from forensic glass evidence: II. *Science & Justice* **1997,** *37* (4), 245-249.
9. Robertson, B.; Vignaux, G. A.; Berger, C. E. H., *Interpreting Evidence: Evaluating Forensic Science in the Courtroom*. 2 ed.; John Wiley & Sons, Ltd: West Sussex, United Kingdom, 2016.
10. Zadora, G.; Martyna, A.; Ramos, D.; Aitken, C., *Statistical Aalysis in Forensic Science: Evidential Value of Multivariate Physicochemical Data*. John Wiley & Sons, Ltd: West Sussex, United Kingdom, 2014.
12. Zadora, G., Evaluation of evidence value of glass fragments by likelihood ratio and Bayesian Network approaches. *Analytica chimica acta* **2009,** *642* (1-2), 279-290.
13. van Es, A.; Wiarda, W.; Hordijk, M.; Alberink, I.; Vergeer, P., Implementation and assessment of a likelihood ratio approach for the evaluation of LA-ICP-MS evidence in forensic glass analysis. *Science & Justice* **2017,** *57* (3), 181-192.
14. Vergeer, P.; van Es, A.; de Jongh, A.; Alberink, I.; Stoel, R., Numerical likelihood ratios outputted by LR systems are often based on extrapolation: When to stop extrapolating? *Science & Justice* **2016,** *56* (6), 482-491.
15. Biedermann, A.; Bozza, S.; Taroni, F., Probabilistic evidential assessment of gunshot residue particle evidence (Part I): likelihood ratio calculation and case pre-assessment using Bayesian networks. *Forensic science international* **2009,** *191* (1-3), 24-35.
16. Biedermann, A.; Bozza, S.; Taroni, F., Probabilistic evidential assessment of gunshot residue particle evidence (Part II): Bayesian parameter estimation for experimental count data. *Forensic science international* **2011,** *206* (1-3), 103-10.

17. Champod, C.; Taroni, F., Bayesian framework for the evaluation of fibre transfer evidence. *Science & Justice* **1997,** *37* (2), 75-83.
18. Causin, V.; Schiavone, S.; Marigo, A.; Carresi, P., Bayesian framework for the evaluation of fiber evidence in a double murder--a case report. *Forensic science international* **2004,** *141* (2-3), 159-70.
19. Collins, A.; Morton, N. E., Likelihood ratios for DNA identification. *Proceedings of the National Academy of Sciences of the United States of America* **1994,** *91*, 6007-6011.
20. Biedermann, A.; Taroni, F., Bayesian networks for evaluating forensic DNA profiling evidence: a review and guide to literature. *Forensic Science International: Genetics* **2012,** *6* (2), 147-157.
21. Curran, J. M.; Buckleton, J. S., An investigation into the performance of methods for adjusting for sampling uncertainty in DNA likelihood ratio calculations. *Forensic Science International: Genetics* **2011,** *5* (5), 512-516.
22. Curran, J. M.; Buckleton, J.; Triggs, C. M., The robustness of a continuous likelihood approach to bayesian analysis of forensic glass evidence. *Forensic science international* **1999,** *104*, 91-103.
23. Bolck, A.; Weyermann, C.; Dujourdy, L.; Esseiva, P.; van den Berg, J., Different likelihood ratio approaches to evaluate the strength of evidence of MDMA tablet comparisons. *Forensic science international* **2009,** *191* (1-3), 42-51.
24. Bolck, A.; Ni, H.; Lopatka, M., Evaluating score- and feature-based likelihood ratio models for multivariate continuous data: applied to forensic MDMA comparison. *Law, Probability and Risk* **2015,** *14* (3), 243-266.
25. Bolck, A.; Alberink, I., Variation in likelihood ratios for forensic evidence evaluation of XTC tablets comparison. *Journal of Chemometrics* **2011,** *25* (1), 41-49.
26. Farmer, N.; Meier-Augenstein, W.; Lucy, D., Stable isotope analysis of white paints and likelihood ratios. *Science & Justice* **2009,** *49* (2), 114-9.
27. Franco-Pedroso, J.; Ramos, D.; Gonzalez-Rodriguez, J., Gaussian Mixture Models of Between-Source Variation for Likelihood Ratio Computation from Multivariate Data. *PLoS One* **2016,** *11* (2), e0149958.
28. Frost, D.; Ishihara, S. In *Likelihood Ratio-based Forensic Voice Comparison on L2 speakers: A Case of Hong Kong native male production of English vowels*, Australasian Language Technology Association Workshop, 2015; pp 39-47.
29. Gonzalez-Rodriguez, J.; Drygajlo, A.; Ramos-Castro, D.; Garcia-Gomar, M.; Ortega-Garcia, J., Robust estimation, interpretation and assessment of likelihood ratios in forensic speaker recognition. *Computer Speech & Language* **2006,** *20* (2-3), 331-355.
30. van Leeuwen, D. A.; Brümmer, N., The distribution of calibrated likelihood-ratios in speaker recognition. *Interspeech* **2013**.
31. Martyna, A.; Sjastad, K.-E.; Grzegorz, Z.; Ramos, D., Analysis of lead isotopic ratios of glass objects with the aim of comparing them for forensic purposes. *Talanta* **2013,** *105*, 158-166.
32. Michalska, A.; Martyna, A.; Ziba-Palus, J.; Zadora, G., Application of a likelihood ratio approach in solving a comparison problem of Raman spectra recorded for blue automotive paints. *Journal of Raman Spectroscopy* **2015,** *46* (9), 772-783.
33. Menzyk, A.; Martyna, A.; Zadora, G., Evidential value of polymeric materials-chemometric tactics for spectral data compression combined with likelihood ratio approach. *The Analyst* **2017,** *142* (20), 3867-3888.
34. Martyna, A.; Zadora, G.; Stanimirova, I.; Ramos, D., Wine authenticity verification as a forensic problem: an application of likelihood ratio test to label verification. *Food chemistry* **2014,** *150*, 287-295.
35. Muehlethaler, C.; Massonnet, G.; Hicks, T., Evaluation of infrared spectra analyses using a likelihood ratio approach: A practical example of spray paint examination. *Science & Justice* **2016,** *56* (2), 61-72.

36. Ramos, D.; Haraksim, R.; Meuwly, D., Likelihood ratio data to report the validation of a forensic fingerprint evaluation method. *Data Brief* **2017,** *10*, 75-92.
37. Latkoczy, C.; Becker, S.; Ducking, M.; Gunther, D.; Hoogewerff, J. A.; Almirall, J. R.; Buscaglia, J.; Dobney, A.; Koons, R. D.; Montero, S.; Peijl, G. J. Q. v. d.; Stoecklein, W. R. S.; Trejos, T.; Watling, J. R.; Zdanowicz, V. S., Development and Evaluation of a Standard Method for the Quantitative Determination of Elements in Float Glass Samples by LA-ICP-MS. *Journal of forensic sciences* **2005,** *50* (6), 1327-1341.
38. Pitts, K.; Trejos, T.; Watling, J. R.; Almirall, J., A guide for the quantitative elemental analysis of glass using laser ablation inductively coupled plasma mass spectrometry. *Atomic Spectroscopy* **2006,** *27* (3), 69-75.
39. Trejos, T.; Koons, R.; Becker, S.; Berman, T.; Buscaglia, J.; Duecking, M.; Eckert-Lumsdon, T.; Ernst, T.; Hanlon, C.; Heydon, A.; Mooney, K.; Nelson, R.; Olsson, K.; Palenik, C.; Pollock, E. C.; Rudell, D.; Ryland, S.; Tarifa, A.; Valadez, M.; Weis, P.; Almirall, J., Cross-validation and evaluation of the performance of methods for the elemental analysis of forensic glass by μ-XRF, ICP-MS, and LA-ICP-MS. *Analytical and bioanalytical chemistry* **2013,** *405* (16), 5393-409.
40. ASTM E2926-13: Standard Test Method for Forensic Comparison of Glass Using Micro X-ray Fluorescence (μ-XRF) Spectrometry. ASTM International: West Conshohocken, PA, 2013.
41. ASTM E2927-13: Standard Test Method for Determination of Trace Elements in Soda-Lime Glass Samples Using Laser Ablation Inductively Coupled Plasma Mass Spectrometry for Forensic Comparisons. ASTM International: West Conshohocken, PA, 2013.
42. ASTM 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 International: West Conshohocken, PA, 2012.
43. Adam, C., *Essential Mathematics and Statistics for Forensic Science*. John Wiley & Sons Ltd.: West Sussex, UK, 2010.
44. Lucy, D. *R Package "comparison": Multivariate likelihood ratio calculation and evaluation*, R Package Version 1.0-4; 2013.
45. Mair, P.; Leeuw, J. D.; Hornik, K. *R Package "isotone": Active Set and Generalized PAVA for Isotone Optimization*, R Package Version 1.1-0; 2015.
46. Mullen, K. M.; Stokkum, I. H. M. v. *R Package "nnls": The Lawson-Hanson algorithm for non-negative least squares (NNLS)*, R Package Version 1.4; 2012.
47. Ramos, D.; Gonzalez-Rodriguez, J., Reliable support: Measuring calibration of likelihood ratios. *Forensic science international* **2013,** *230* (1-3), 156-169.
48. Brümmer, N.; du Preez, J., Application-independent evaluation of speaker detection. *Computer Speech & Language* **2006,** *20* (2-3), 230-275.
49. Zadora, G.; Ramos, D., Evaluation of glass samples for forensic purposes — An application of likelihood ratios and an information–theoretical approach. *Chemometrics and Intelligent Laboratory Systems* **2010,** *102* (2), 63-83.
50. Ramos, D.; Gonzalez-Rodriguez, J.; Zadora, G.; Aitken, C., Information-theoretical assessment of the performance of likelihood ratio computation methods. *Journal of forensic sciences* **2013,** *58* (6), 1503-18.
51. Brümmer, N.; du Preez, J., Application-independent evaluation of speaker detection. *Computer Speech & Language* **2006,** *20* (2-3), 230-275.
52. Meuwly, D.; Ramos, D.; Haraksim, R., A guideline for the validation of likelihood ratio methods used for forensic evidence evaluation. *Forensic science international* **2017,** *276*, 142-153.