J Forensic Sci, Nov. 1999, Vol. 44, No. 6

Authors' Response

Sir:

One of our intended purposes in writing the referenced article, in addition to presenting our research results, was to stimulate discussion among forensic scientists concerning the important topic of the use of statistics in evaluating items of trace evidence. We wish to thank Curran et al. for initiating this discussion and giving us the opportunity to clarify and expand upon a few points that we made in our original paper. The letter writers indicate that our aim was to show that statistics are "pointless". Nothing could be further from the truth. In fact, we are proponents of the appropriate and correct use of statistics in the evaluation of evidence. However, we do not advocate the calculation of purportedly exact statistical measures that may be interpreted without consideration of the underlying uncertainties. We wish the readers to recognize the difficulty, if not impossibility, of calculating frequency of occurrence statistics when using highly discriminating analytical techniques to evaluate evidence whose characteristics vary over both location and time. Although our paper is concerned with the elemental analysis of glass, similar considerations apply when evaluating many items of trace evidence using well-accepted methodologies. The following comments address specific points raised by Curran et al.

The discriminatory power of a technique is not only interesting, but it is also quite useful to scientists making decisions whether or not to use the technique. It should also be of interest to triers of fact when considering what significance to place on analytical results in legal proceedings. In our paper, we do not use or explicitly calculate "discriminatory power" (a term used by the letter writers). However, the data we present clearly indicates the high degree of discrimination among glass sources obtained using a combination of refractive index (RI) and elemental analysis. The link between RI and elemental composition is not significant in this work. The RI of a glass fragment is a direct result of both its total chemical composition and thermal history, and it is independent of any single element concentration. There is no need to consider elemental analysis conditional upon RI as suggested by the letter writers. In fact, our study, which consists of all evidentiary glass for which the FBI Laboratory obtained triplicate analyses from 1990 to 1996, includes no two sources with the same elemental composition, regardless of RI.

The concept of information content is a valid measure of discrimination, within the context that it is used in our paper. The information content, as we defined it, is a measure of the maximum number of distinguishable sources that could possibly exist within the compositional range exhibited in a set of samples. It is a useful measure of the relative discrimination capability of a given technique and serves as a benchmark for comparison of alternate techniques for a given analysis. For example, several forensic laboratories are currently considering the use of ICP-MS instead of ICP-AES for compositional analysis of glass. The question of whether one obtains better discrimination capability by determining 30 elements with relative standard deviations (RSDs) in the 10–50% range by ICP-MS or the 10 elements with 1–5% RSDs by ICP-AES can be answered by comparing the information content of the two methods. As we pointed out in our paper, the information content provides no information about the distribution of glass specimens within the elemental and RI combinations. Despite Curran et al.'s concern that information content be given any credence at all, this measure (although calculated differently than we defined it in our paper) has been widely used and has stood as a landmark concept in information theory as applied to analytical spectroscopy for over 20 years (1).

We agree that we have answered Curran et al.'s pre-data question and not their post-data question. The purpose of our article was to demonstrate that the analytical method used provides information that can be used for excellent source discrimination—a pre-data question. The post-data question as posed by the authors is applicable to evaluation of evidence in a case framework, a situation not addressed in our article. We agree that their post-data question is best answered by a Bayesian approach, precisely because the question is framed within that approach. We think it important to note, however, that the discrimination capability of the analytical method is an intrinsic part of the calculation of the likelihood ratio and any assessment of the significance of the evidence. The Bayesian approach is one method of assessing the significance of a finding of indistinguishability between glass fragments recovered from a suspect and those from a broken glass object. In simple cases, where probability distributions for all measured parameters in the appropriate crime scene and alternate hypothesis environments and transfer and persistence parameters are known, the Bayesian approach may be viable. Additionally, we never stated that glass databases are not the most reliable way of assessing the value of evidence. We agree that *appropriate* databases are the best way of calculating frequency of occurrence statistics. However, we state again that, when using highly discriminating analytical methods and considering items of evidence whose distributions vary over both location and time, it may not be possible to obtain the databases needed for the Bayesian approach. Application of any statistical approach to probability calculations when population distributions are unknown is dangerous and may produce misleading results.

A major portion of the letter consists of a primer on the calculation of likelihood ratios, which is a summary of the authors' work in this area. We suggest that interested readers read the original articles (2,3) for a full derivation of the equations used in the letter. What is not mentioned in the letter and the authors' other articles is the uncertainty associated with each term in their equations. In the denominator of their first equation, the probability of the evidence given no contact depends upon having a database of glass from wherever the defendant's alibi may be. The values, which must be used for the transfer and persistence terms, are highly subjective and subject to order of magnitude errors in realistic case situations (see Reference 2 for examples). The quantity lr_{cont} is an interesting approach, particularly when coupled with the use of Hotelling's T^2 for multivariate data. However, as pointed out by Curran et al., lr_{cont} is roughly proportional to 1/P. An important point that we have made in our paper is that the value of P is extremely small. One can dispute the details of the calculation of the rarity of a particular glass, but it is indisputable that as more discriminating methods of analysis are used, the probability of two different sources of glass being indistinguishable decreases and the likelihood ratio increases. Our comment concerning likelihood ratio calculations, to which Curran et al. seem to have such a strong objection, is that the number cannot be calculated with any degree of precision. However, this is unimportant if an analytical method is used that assures that the number is so large as to be highly significant for indistinguishable specimens.

We agree that there is some error associated with each measurement in glass. In fact, there is measurement error associated with any analytical measurement in any field of endeavor. There is a vast field of chemistry literature detailing non-Bayesian methods

of dealing with analytical error. The discrimination potential of a method is determined by the magnitude of the measurement error plus sample heterogeneity relative to the range across similar samples. The fact that measurement error exists does not "prove once again that the Bayesian approach is necessary". The equations of likelihood ratio are an interesting academic exercise and provide a framework for qualitative consideration of the factors involved in assessing the significance of matching analytical data. We appreciate the discussion of this method and leave it to the readers of this journal to determine whether the calculation of likelihood ratios is a reasonable and legally acceptable approach for presentation of evidence to a court of law.

We do not feel that we used a "fixed bin" approach inappropriately here. Our bins are not of fixed width, a point which we discussed in detail in our paper. The selection of bin widths based on measurement precisions is an appropriate method for comparison of specimens of similar compositions. The bin means are fixed in our calculations of our measures of information content and most common composition. However, as we state in our article, if we were to use our data to calculate the frequency of occurrence for an evidentiary specimen, we would use a floating bin for each variable with a position and width based on the analytical mean and standard deviation calculated from replicate samples of the evidentiary specimen. The justification of 0.0002 as a bin width for RI is curious, we agree. Rightly or wrongly, however, it is a number that has been widely used as a fixed cutoff for source differentiation by many glass examiners for roughly 20 years (4). RI differs from other parameters in our study, in that RI measurement uncertainty (bin width) does not vary with RI measurements (bin center locations). Therefore, we chose 0.0002 as a constant bin width, recognizing that it is smaller than the 12σ widths of the other bins. For purposes of casework assessment of glass fragments, we agree that a fixed cutoff of 0.0002 is generally inappropriate and it is preferable to use a statistical test criterion based on repeated measures of the glass fragments in question. Curran et al.'s reference to other published RI density distributions is curious, as these are clearly inappropriate for case-specific situations. For example, a frequency distribution given in a 1978 article about glass in England is certainly not applicable to a 1999 case in the United States. However, this does not matter for our approach, since the choice of element concentration bin widths is based on analytical precision and source heterogeneity and has nothing to do with probability density distributions. The selection of 12σ bin widths for element concentrations is explained in our paper. That the bins are wide is supported by the fact that two specimens having data at adjacent bin centers are clearly distinguishable by any reasonable statistical test. In fact, two specimens lying near opposite edges within the same bin are readily distinguishable using the match criteria of the FBI Laboratory. The use of a calculated standard deviation measure in setting bin widths does imply some degree of normal distribution to the underlying data. We agree that for some broken glass objects, some or all of the measured parameters do not exhibit a normal distribution. This should have no affect on our selection of bin widths for purposes of assessing the variations in observed compositions. It would, however, adversely affect the commonly used methods of statistical evaluation of the data (i.e., pooled t-test, calculations of LR) from that broken object.

The letter writers state that standard deviations are unknown for our specimens. In fact, as pointed out repeatedly in our paper, the standard deviations are calculated based on measurements from triplicate fragments from each specimen. That the standard deviation is not constant across samples is the point of our Fig. 1 and the related selection of variable bin widths. The standard deviation is unknown only in the sense that three samples may not be enough to calculate a standard deviation when the distribution is not normal. Generally, a t-test of means is appropriate for comparison of specimens, because, as shown clearly in our Fig. 1, two specimens with similar means will have similar standard deviations. Comparison of two samples with dissimilar means, where the standard deviations are different, is a trivial exercise because widely different means are readily distinguished by any statistical test.

Curran et al. make several comments concerning the state of casework samples and the assumption that our data were obtained from "perfect" samples. All of the samples in this study were derived from casework samples, either as specimens of known broken windows, fragments recovered from clothing and other sources, or comparison exemplars, such as alibi sources. Approximately one-fourth of our specimens were recovered fragments and three-fourths were from known broken glass objects. Samples were cleaned with concentrated nitric acid prior to analysis (5), a procedure that removes contamination and results in consistent element concentration measurements. Whether or not questioned samples exhibit a preponderance of fragments containing an original surface is a moot point. No one has reported and we have seen no evidence of measurable differences between the concentrations of the measured elements in surface and bulk samples of cleaned glass. The claim that the samples are too small is not true, in that they all meet the size requirements for elemental analysis according to the FBI Laboratory protocols in effect at the time of their examination. The claim that this "error" is potentially serious is untrue, because no error of the type Curran et al. describe exists. The comment that because the samples were weighed, they are atypical of recovered glass fragments reveals a lack of analytical experience of the letter writers. Samples as small as 100 μg are routinely weighed and analyzed in many analytical laboratories. Microbalances are capable of weighing samples with a precision of 0.1 μg, which equates to a relative precision of 0.1% for a 100 μg fragment.

We do not understand why the processing method of the data in our Fig. 1 is unclear, since it is described in the text. Figure 1 is a plot for each element of the RSD of the triplicate samples for each specimen versus the mean for that specimen. To convert this data to bin widths, a smooth curve was drawn through the points and the value for each 12σ bin width was calculated by multiplying the standard deviation value corresponding to the mean concentration at each bin center by 12. The comments about serious data editing to eliminate duplicate samples seem unwarranted to us. Since these are case-derived samples, many of which are of unknown ultimate source, we limited the number of samples to include equal weighting (3 replicates) from each source. The number of samples was not reduced from 1504 to 204, as Curran et al. suggest. Rather, removal of samples with less than three replicates and those duplicate samples from the same case reduced the data set from 1504 samples to 612 (triplicate samples from 204 specimens). Limiting the number of samples in this way will not diminish the correlation coefficients, but rather would increase them. For example, if we were to include 100 samples from the same source it would generate a symmetric cluster of points about a mean value, the size of the cluster dictated by the combined analytical precision and sample variation. Such a cluster of points in a regression plot would lessen the value of any calculated correlation coefficient.

We do not agree with the comment that casework samples are an odd set that is not as useful as samples collected from persons not

associated with crime. There is no evidence we can find in the literature or in our considerable past experience that there is any difference in the distribution of any of the measured parameters between glass recovered from people suspected of crimes and from those not suspected of being associated with crime. The data set of glass from people unassociated with crime would be an interesting one for comparison with other existing databases. However, such a database does not exist, because it would be impossible to collect. In various comments throughout their letter, Curran et al. suggest that to interpret our data in a Bayesian context we would need something on the order of 10^{11} specimens collected from random individuals unassociated with crime. Further, we would need perhaps 10 analyses of each specimen to correctly assess standard deviations, normality of parameter distributions, and to use multivariate versions of the t-test, such as Hotelling's T^2. Collection of such a database is impossible because people unassociated with crime involving broken glass typically do not have many fragments from the same source on their persons (6).

The calculations of coefficients of linear regression were based upon raw data, not binned data, because binning first would have decreased the information content of the data, as stated by Curran et al. We apologize for not making this clearer in the text of the article. The caption of Fig. 3 should have read, "The distribution of Al and Mn among glass specimens." The statement that the skewed distributions shown in our Figs. 1 and 2 would result in nonlinear correlations between pairs of variables is not correct. Figure 1 displays precision of measurements, which effectively has no bearing on correlation coefficients. The fact that samples are not evenly distributed across Fig. 2 cannot be directly translated into correlation coefficients, because it cannot be discerned from Fig. 2 which point in one element plot corresponds with a point in another element plot. Thus nothing can be said about the linearity of correlations by observing Fig. 2, despite the claim of Curran et al. The elements Al and Mn were selected for the scatter plot shown as our Fig. 3 because this is the pair of variables with the *best* correlation. No nonlinear relationships are apparent from visual observation of this figure or similar figures of every other pairwise combination of variables. In summary, we find no evidence of strong correlations between pairs of variables, either linearly or nonlinearly. Curran et al. use the fact that we observe no correlation between RI and composition as evidence of our inability to detect correlations between variables. In fact, there should be no direct correlation between RI and the concentrations of any single element. The sum of all measured elements in our analytical protocol is roughly 18% of the total mass of the glass fragment. The RI is more profoundly influenced by the elements not determined in our protocol (such as silicon, lithium, potassium, and lead) than by the elements determined. If we had measured the concentration of every element and the RI for each sample, then two dimensions of redundancy would exist in our data (the sum of all oxides must be 100% and the RI is roughly calculable from the composition). At any rate, the lack of correlation is not caused by the "serious data editing" which Curran et al. purport to exist in our database.

The comment concerning the inability to prove lack of dependence among 11 variables is a good point. It is possible that there is an interdependence of element concentrations such that the data could be rotated in 11-dimensional space to form linear combinations of variables without *significant* loss of discrimination among samples. If such a dependence exists, then multiplying probabilities together as we did would result in some overestimation of discrimination capability. We have seen no indication of this intervariable correlation, but because it is possible, we suggest that our calculations give reasonable, but not exact estimates of the probability of matches among randomly collected glass fragments. As Curran et al. point out, the "curse of dimensionality" is a consideration in using multivariate databases. The number of samples required to form a probability density function in 11 dimensions is unrealistically large. Variable reduction, such as by factor analysis, can be used to reduce the dimensionality to facilitate classification decisions and for convenient plotting of results. However, any variable reduction method results in loss of information. In the comparison of evidentiary specimens, it is of paramount importance to avoid false associations in that these could lead to incorrect consequences for an innocent accused. Therefore, all measured variables must be indistinguishable to result in a conclusion of two fragments of glass (or hair, soil, fibers, or any other transfer evidence) having come from a single source. Reduction of dimensionality to make the data fit a simple statistical model for purposes of calculating probability statistics does not justify the loss of information and consequent increase in the number of false associations. Another practical consequence of variable reduction methods is that the new factors formed by linear combinations of variables do not have readily discernable physical sense. That is, a factor that is a linear combination of 10 element concentrations and RI cannot be explained to the participants of a criminal proceeding in a manner that they will understand and whose significance they will appreciate.

In summary, we are not against the use of statistics in the evaluation of forensic evidence. Rather, we are proponents of it and believe that the Bayesian approach has considerable merit in the appropriate applications. However, we are stronger advocates of the use of good analytical methods to provide accurate, precise analytical data with as much discrimination capability as possible. The statistical evaluation of such data is much more difficult, particularly for manufactured items such as glass, than it is for data such as RI alone. It is apparent without calculating any probability statistics that a chance matching of randomly selected samples is extremely small and, as a result, the exact calculation of statistic figures is not important to the trier of fact. The Bayesian approach is useful in that factors other than population frequency can be considered in evaluating the significance of the evidence under several alternate hypotheses. We believe that it is more important to use highly discriminating and reliable analytical methods, even if they cannot be used to calculate an exact probability number, than it is to use poorer analytical methods or data reduction in order that statistics can be calculated.

References

1. Kaiser H. Foundations for the critical discussion of analytical methods. Spectrochim Acta B 1978;33B:551–76.
2. Walsh KAJ, Buckleton JS, Triggs CM. A practical example of glass interpretation. Sci Just 1996;36:213–8.
3. Curran JM, Triggs CM, Almirall JR, Buckleton JS, Walsh KAJ. The interpretation of elemental composition measurements from forensic glass evidence II. Sci Just 1997;37:245–9.
4. Miller E. Forensic glass comparisons. In: Saferstein R, editor. Forensic science handbook. Englewood Cliffs (NJ): Prentice-Hall 1982;139–83.
5. Koons RD, Fiedler C, Rawalt, RC. Classification and discrimination of sheet and container glasses by inductively coupled plasma-atomic emission spectrometry and pattern recognition. J Forensic Sci 1988;33: 49–67.
6. Lau L, Beveridge AD, et al. The frequency of occurrence of paint and glass on the clothing of high school students. Can Soc Forensic Sci J 1997;30:233–40.

Robert D. Koons, Ph.D.
JoAnn Buscaglia, Ph.D.
Forensic Science Research Unit
FBI Academy
Quantico, VA 22135