

## Interpretation of Glass Composition Measurements: The Effects of Match Criteria on Discrimination Capability\*

**REFERENCE:** Koons RD, Buscaglia J. Interpretation of glass composition measurements: the effects of match criteria on discrimination capability. *J Forensic Sci* 2002;47(3):505–512.

**ABSTRACT:** The concentrations of ten elements in 209 unrelated glass specimens received as evidence were used to assess the frequencies of errors of false association (Type II errors) using several comparison criteria at specified significance levels (Type I errors). Pairwise comparisons of the samples using either the equal-variance t-test or Welch's modification (unequal variances) result in a small number of errors of false association, even when adjusting the significance level (Bonferroni correction) for multivariate comparisons. At the 95% confidence level (overall Type I error of 0.05, or individual element comparison error of 0.005), only two Type II errors are made in 21736 comparisons (0.009%) when using the equal-variance t-test for comparison of sample means. In this study, the range overlap test using three replicate measurements per specimen results in no errors of false association. Most specimen pairs in this data set are readily discriminated either by differences in the concentrations of several elements or by an extremely large difference in the concentrations of one or more element.

**KEYWORDS:** forensic science, criminalistics, trace evidence, glass, elemental analysis, statistics, ICP-AES

The forensic comparison of recovered glass fragments with specific broken objects is a well-accepted trace evidence examination. A variety of optical, physical, and chemical properties of glass fragments has been successfully utilized for these comparisons. Despite the number of possible comparison parameters, the majority of forensic glass examiners have consistently relied primarily on refractive index and density measurements. There are several reasons for the frequent use of these properties. They have been shown to possess relatively good source discrimination capabilities; well-documented, nondestructive analytical methods are widely available; and there is a long history of their acceptance in most legal systems. As a result, there have been many studies concerning both the interpretation and significance of comparisons using refractive index and density measurements. A large body of literature exists that discusses various statistically-based approaches that can be used to place some measure of significance on the finding that two or more glass fragments are analytically indistinguish-

able. Nearly all of these published studies are concerned with refractive index measurements. Although there have been fewer studies concerning their interpretation, density measurements have been accepted as valid, in large part, by analogy to refractive index measurements. These studies concerning methods of significance testing, particularly those using refractive index measurements, have recently been summarized by Curran et al. (1).

Although RI and density are widely used for comparison of forensic evidence and offer relatively good source discrimination capability, their discrimination potential is limited in comparison with that of chemical composition. The very high discrimination capabilities of methods providing precise measures of elemental composition have been well documented (2). However, most of the literature concerning elemental analysis of glass describes the development and validation of analytical procedures rather than evaluation of forensic significance. Initial approaches based on neutron activation analysis, X-ray fluorescence spectroscopy, and optical emission spectrometry have been replaced for the most part by measurements of element concentrations by inductively coupled plasma-atomic emission spectroscopy (ICP-AES) and, most recently, inductively coupled plasma-mass spectrometry (ICP-MS). Limited availability of instrumentation in crime laboratories and rapid changes in analytical methods have hampered adoption of consistent analytical protocols, including lists of analyte elements. Consequently, match criteria for elemental measurements have not been as well documented as those for refractive index, or even density measurements.

The various comparison criteria applied in trace evidence evaluations have historically relied on traditional hypothesis tests, such as the Student's t-test. The very nature of these test statistics poses an interesting problem in their application to evidentiary comparisons. Hypothesis tests are performed to specified confidence levels, which define the percentage of Type I errors that can be expected for the comparison of mean values of measured variables in two sample populations. Type I errors are defined as errors of failing to accept the null hypothesis when it is true. In this context, a Type I error is forming the erroneous conclusion that there is a significant difference between the compared specimens when, in fact, there is not. Type I errors are of less concern in a forensic context than are Type II errors, failing to reject the null hypothesis when it is false. A Type II error, or false inclusion, is considered the more egregious error in forensic comparisons, because it may incriminate a truly innocent subject. Unfortunately, conventional hypothesis tests do not offer a means of directly specifying an acceptable level of Type II errors. Further, there is an inverse relationship between Type I and II errors, so that adjustments to test criteria to lower the frequency of Type I errors result in an increase in the

<sup>1</sup> Research Chemist, Forensic Science Research Unit, FBI Laboratory, FBI Academy, Quantico, VA.

\* This is Publication No. 01-20 of the Laboratory Division of the Federal Bureau of Investigation (FBI). Names of commercial manufacturers are provided for information only, and inclusion does not imply endorsement by the FBI.

Received 11 Oct. 2001; and in revised form 3 Dec. 2001; accepted 3 Dec. 2001.

number of Type II errors. There is always a compromise between these two types of errors, with only the level of Type I errors readily predicted from the comparison criteria. This paper reports the results of a study in which a database containing analytical results from unrelated specimens was used to assess the Type II error rate at given Type I error levels.

Most of the studies concerning the effects of various comparison criteria and the significance of indistinguishability of compared glass fragments are based on refractive index measurements and, therefore, nearly all of them utilize univariate test statistics. However, interpretation of elemental concentration data is complicated by the fact that one must consider multiple comparisons. For example, when two samples from the same source are compared using hypothesis testing with a single variable at a particular significance level, say  $\alpha = 0.05$ , then there is a 5% probability of falsely excluding them. Further comparison using a second variable will produce an additional 5% false exclusion rate. For multivariate comparisons, the overall rate of Type I errors is approximately equal to  $n \times \alpha$ , where  $n$  is the number of variables. For comparisons involving ten variables, as is typical with elemental concentration measurements of glass, an unacceptably high rate of Type I errors results unless a correction is performed. Approaches used to address this problem include reducing the  $\alpha$  level (the Bonferroni correction), using a multivariate form of the comparison test, or using the continuous likelihood ratio approach. Curran et al. state that lowering the  $\alpha$  level to  $\alpha/n$  will result in extremely high levels of Type II errors (3). Instead, they recommend the use of Hotelling's  $T^2$ , a multivariate form of the  $t$ -test, and continuous likelihood-based approaches for elemental comparisons (4). However, it should be noted that each of these methods has limitations. The Hotelling's  $T^2$  test requires too many replicate measurements to be practical for use in forensic case work. For example, the comparison of ten variables requires the analysis of at least twelve replicate samples. The continuous approach requires knowledge of multidimensional probability density functions that are, for the most part, nonexistent and impossible to construct.

Another method of reducing the overall Type I error rate in multivariate comparisons is to reduce the number of variables. Once again, a compromise must be made. The use of too many variables increases the probability of making a Type I error, whereas using too few variables limits the discrimination capability. The optimum balance will depend upon sample constraints, analytical figures of merit, interdependence of variables, and the distribution of measured values within glass objects and across all sources. For example, the addition of elements to an analytical scheme may not improve discrimination if they are correlated with other elements, have poor precision within a single glass object, or exhibit a narrow range of values among glass sources. On the other hand, elimination of an element from an analytical protocol may fail to reveal a difference between two sources that vary greatly in the concentrations of that element.

In this study, a test data set of the concentrations of ten elements in evidentiary glass specimens, determined by ICP-AES, was used to assess the frequency of errors of false associations. Specifically, the effects of applying the Bonferroni correction and decreasing the number of variables, and the relative discrimination capabilities of the elements were evaluated.

## Methods

### *Specimen Selection*

The glass specimens selected for this study were received as evidence by the FBI Laboratory during the period of 1990–2000. For

those cases where it is appropriate, according to the FBI Laboratory protocol, the concentrations of up to ten elements are determined by ICP-AES in glass fragments. Three fragments from each specimen are analyzed separately to provide a measure of the variability of each measured element concentration within the source object. However, in some instances the limited size of the evidentiary specimen requires that the triplicate samples must be taken from a single fragment. The FBI Laboratory compiles the results of these analyses as part of its experience database. At the time this study was undertaken, the results of elemental analysis of 1806 analytical samples were contained in the database.

For this study, the full database was reduced in size to make a test sample set suitable for the determination of the rates of matching errors. Only those evidentiary specimens for which three fragments had been analyzed for the full ten-element suite were included in the test set. For each case, one specimen was initially selected and its data placed into the test sample set. The results for every other specimen in that case were compared to the selected data. If a second specimen in that case were clearly analytically distinguishable in one or more parameter, it was considered to represent a second source of glass, and it was also selected for the test sample set. This process was continued for all specimens, comparing each one with all previously selected specimens from that same case, until a group was selected representing the distinguishable separate sources of glass received in each particular case. This process was repeated for each case in the original database, and all selected data were compiled together into the test set. This method of sample selection ensures that the test set does not include data from two specimens from the same broken object in a given case. However, no cross-case comparisons were considered in the selection of samples. The test set in this study, therefore, represents an unbiased sample, in the sense that each glass specimen is represented only one time. The resulting data set used for this study consists of 209 specimens from 148 cases, each analyzed in triplicate, or 627 analytical samples. This data set is similar to that used in a previous publication (2). However, new specimens that were analyzed after the earlier study was completed have been added to this data set. A few specimens for which an element concentration was missing and nonevidentiary research specimens that were included in the prior study were excluded. This test set should be a realistic representation of an appropriate population suitable for frequency-of-occurrence studies because it consists of only those seemingly unrelated glass specimens submitted as evidence.

### *Methods of Analysis*

All samples in the test data were analyzed using well-established methods in use in the FBI Laboratory for examination of evidentiary specimens. The concentrations of the ten elements, Al, Ba, Ca, Fe, Mg, Mn, Na, Sr, Ti, and Zr, were determined in 2–8 mg samples using ICP-AES. Details of the cleaning, dissolution, and analysis procedures are presented elsewhere (5,6). One or more well-characterized standard reference glasses (NIST, Gaithersburg, MD) were prepared and analyzed in the same manner as the samples with each case. The accuracy of the results for these reference glasses was verified for each element before sample results for that case were added to the experience base, from which the data for this study were drawn. This continuing check on the accuracy of the results allows the data to be used for studying the distribution of the measured parameters in the glass population represented by the experience base. The precision, as measured by the variation of the triplicate analytical results for each specimen, reflects both

the analytical reproducibility and the source variability of the measured glass object.

Refractive indices (RI) of some specimens were determined using the Emmons double-variation method (7). According to the FBI Laboratory protocol for the Emmons double-variation method, RI measurements are performed on multiple analytical samples, but only a single RI value to the fourth decimal place is recorded in the experience base. Since no precision measure is recorded with each RI value, this variable could not be included in this statistical study. For a period of time, RI measurements from known-source objects only were placed into the RI experience base. Therefore, RI data are available for only 151 of the 209 specimens in the test set.

**Results and Discussion**

The principal goal of this study was to evaluate the relationship between Type I and Type II errors in glass comparisons using the elemental composition data. This relationship was studied by making all pairwise comparisons of the 209 specimens (21736 comparisons) for each of the ten elements using the Student’s t-test as a test criterion. Several  $\alpha$  levels were selected to define specified levels of Type I errors. The test set was constructed such that every specimen represents a distinct source. Therefore, at each  $\alpha$  level, the number of Type II errors can be directly measured, because it is equal to the number of indistinguishable specimen pairs.

When using the t-test to compare the mean values for two samples, assumptions concerning the shape and variance of each distribution are made. Use of the t-test assumes that each of the sample populations being compared is normally distributed. For glasses formed from well-mixed solutions, such as in high-volume float and bottle production, the element variation is random and measured concentrations can reasonably be expected to be normally distributed, particularly for those elements present at measurable levels using ICP-AES. Regardless, it has been reported that minor deviations from normality do not adversely affect the use of the t-test (8,9). A second consideration for use of the t-test is whether the variances of the two sample populations being compared are equal. It has been shown that element concentrations in two glass specimens having similar mean concentrations generally have equal variances (2). This similarity of variances results from the fact that instrument precision is typically a more important contributor to the overall variance than is the within-source heterogeneity. When the mean concentrations of an element differ significantly in two specimens, the variances will also differ, because instrument precision varies with concentration. However, such specimens are readily excluded as having come from the same source by any comparison method, including casual observation. For this study, t-tests using both equal and unequal variances (Welch’s modification) for the compared sample populations were applied to all pairwise combinations of the data set. For the equal-variance t-test, a single pooled standard deviation value for the two specimens in each pair was used and, for the unequal-variance test, the Welch modification was used for each comparison, regardless of whether there was a significant difference between the two standard deviations. All t-tests were evaluated using 2-tailed test criteria.

The results of pairwise comparisons using the t-test with both equal and unequal variances at five probability levels are shown in Table 1. The  $\alpha$  levels for each individual comparison and the corresponding overall Type I error rates for comparison of ten elements are presented as the upper- and lower-column headings, respectively. As shown in Table 1, when using the t-test to compare

glass specimens with ten element concentrations as variables, the number of Type II errors is very small. At a relatively high percentage of over all Type I errors (10%), the equal-variance t-test results in no Type II errors, and the unequal-variance t-test results in two Type II errors out of 21736 pairwise comparisons. Using a more conventional overall  $\alpha$  level of approximately 0.05, the occurrence of Type II errors increases to 1/11000 (equal variance) and 1/1800 (unequal variance.) Even when using the relatively wide comparison criteria of 0.01 for overall  $\alpha$ , the frequencies of false inclusions are approximately 1/1000 (equal variance) and 1/130 (unequal variance). As a point of perspective, for two specimens to be barely distinguishable by the equal-variance t-test at the 0.01 overall  $\alpha$  level (individual  $\alpha = 0.001$ ), their mean concentrations of at least one element must differ by greater than 8.6 times the pooled standard deviation. This match criterion is quite wide compared to the two or three standard deviation criteria frequently used in forensic comparisons.

The inverse relationships between Type I and Type II errors for both the equal- and unequal-variance t-tests based on the data in Table 1 are expressed graphically in Fig. 1. The higher percent of Type II errors with the unequal-variance t-test compared to the equal-variance t-test is readily apparent, particularly as the overall  $\alpha$  level is minimized. A glass examiner can use a plot such as the

TABLE 1—Number and percent of indistinguishable pairs using the concentrations of 10 elements.

$\alpha$ Level for Each Test	0.01	0.005	0.002	0.001	0.0005
Indistinguishable Pairs					
<u>Equal Variance</u>					
Number	0	2	7	22	40
Percent	0%	0.009%	0.032%	0.101%	0.184%
<u>Unequal Variance</u>					
Number	2	12	48	168	427
Percent	0.009%	0.055%	0.221%	0.773%	1.96%
Overall $\alpha$ level	0.096	0.049	0.020	0.010	0.0050

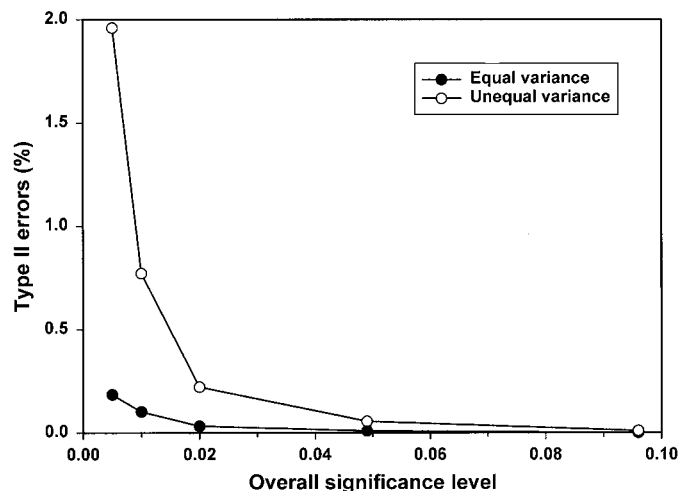


FIG. 1—Frequencies of Type I and Type II errors resulting from 21736 pairwise comparisons of evidentiary glass specimens using t-tests with equal and unequal variances.

one shown in Fig. 1 to select an appropriate test criterion to perform the t-test at predetermined error rates for Type I and Type II errors. The significant conclusion gained from this study is that the Type II error rates are extremely low at acceptable overall Type I error rates. The range of compositions seen in the evidentiary glass specimens in this study is so wide that the likelihood of errors of false inclusion is relatively low, regardless of the match criteria applied.

The summary of the t-test results shown in Table 1 does not provide any information about the number of elements that differ for each distinguishable pair. Such information is useful in understanding the discrimination capability of the analytical method. In evidentiary comparisons, an examiner would want to know not only that two specimens can be distinguished, but also the number of elements that differ and the magnitude of the differences. Any significant difference between two specimens indicates that they originated from different sources. However, if the concentrations of many elements differ between two specimens, the likelihood that they came from two separate sources is greater than if only a single difference existed. The importance of the selection of match criteria used for forensic comparisons is less when multiple element concentrations differ between specimens than when only a single difference exists. In addition, when the  $\alpha$  level of a test is decreased, the increase in the likelihood of making a Type II error is less when the sample means differ in more than one element. Figure 2 illustrates the number of times a specimen pair in this data set can be differentiated by each of the given number of elements using equal- and unequal-variance t-tests. At each probability level studied, the number of specimen pairs is plotted as a function of the number of distinguishable element concentrations. As shown, for most of the specimen pairs, there are five or more distinguishable element concentrations when comparisons are performed at individual test  $\alpha$  levels of 0.001 to 0.005. In other words, most specimen pairs taken from the random sources of glass in this study are very different in their elemental composition patterns as determined by ICP-AES measurements.

An interesting difference in results for the equal- and unequal-variance t-tests appears in Fig. 2. As the  $\alpha$  level for each test using unequal variance is decreased (i.e., fewer Type I errors), the curves shown in the lower portion of Fig. 2 shift to lower numbers of elements having distinguishable concentrations. That is, as match criteria are made wider, as would be done using the Bonferroni correction, for example, there is a decrease in the number of elements with distinguishable concentrations between a pair of specimens from different sources. In contrast, as shown in the upper portion of Fig. 2, the distribution of the number of elements distinguished in pairs of specimens is relatively unaffected by significance level when using the t-test with equal variance. Using this test, most of the pairs are distinguishable by seven or more elements at all  $\alpha$  levels. If the concentrations of each element in two specimens being compared have equal variances, then the equal-variance t-test is the correct test to use and is more powerful for detecting true differences. Whether using equal- or unequal-variance tests, one can lower the  $\alpha$  level and still maintain a high degree of confidence that two separate sources will be distinguishable by several elements using results determined by the ICP-AES method.

Another consideration, not shown in Fig. 2 is that, for most specimen pairs, the magnitudes of the differences in concentrations of one or more element concentrations are very large. For example, as shown in Table 1, using the t-test with equal variances at the  $\alpha = 0.0005$  level, all but 40 of the 21736 specimen pairs (99.8%) are distinguishable by concentration differences in one or more element. The critical value of  $t$  at  $\alpha = 0.0005$  is 10.305. In other

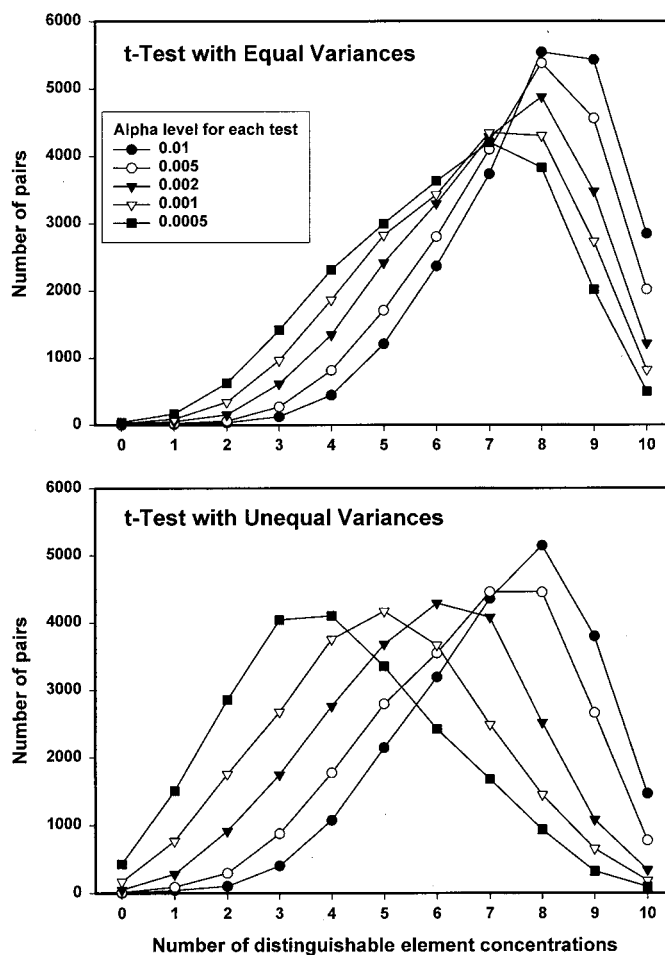


FIG. 2—Number of elements distinguishing 21736 pairs of glass specimens as a function of significance level for equal- and unequal-variance t-tests.

words, 99.8% of the specimen pairs have concentration differences for at least one element that are greater than 10.305 times their pooled standard deviation—a difference that is so large as to be readily apparent with any hypothesis test, including simple observation of the analytical data.

Figure 2 also does not indicate which element concentrations differ between pairs, but rather the total number of differences considering all of the ten element concentrations. The discrimination capability of each individual element can be evaluated by determining the number of differences between specimen pairs by element. The number and percent of distinguishable pairs for each element at the individual  $\alpha$  level of 0.005 using equal- and unequal-variance t-tests are shown in Table 2. For these tests, nearly every element is capable of discriminating 50% or more of the specimen pairs, with most elements having approximately 80% discrimination capability. Strontium is the most discriminating element, with greater than 90% discrimination capability, followed closely by Fe, Al, Ba, and Ti. It should be noted that the specimens in this data set are predominantly flat glass, because of its prevalence in case work. A different order of discrimination capability of the elements would be seen for other glass types. For example, if the data set consisted of an equal mix of glasses having Mg concentrations of approximately 0.2% (typical of containers) and 2.0% (typical of float sheets), then the number of specimen pairs discriminated by Mg alone would be much greater

TABLE 2—Individual element discrimination capability—number and percent of distinguishable pairs for each element at individual  $\alpha = 0.005$ .

	Ca	Fe	Al	Mn	Sr	Mg	Ba	Ti	Zr	Na
Individual $\alpha = 0.005$										
<u>Equal Variance</u>										
No. distinguishable pairs	15321	19096	18495	15324	20385	11115	18841	17673	14340	11909
% Discriminated	70.5%	87.9%	85.1%	70.5%	93.8%	51.1%	86.7%	81.3%	66.0%	54.8%
<u>Unequal Variance</u>										
No. distinguishable pairs	12880	17812	17360	13440	19601	8942	17404	15827	11767	9155
% Discriminated	59.3%	81.9%	79.9%	61.8%	90.2%	41.1%	80.1%	72.8%	54.1%	42.1%

than 50%. For such a data set, Mg would probably occupy a higher position in the order of discrimination capability of the elements than it does in the set used for this study.

Several researchers have suggested that the concentrations of five elements are enough to provide adequate glass source discrimination capability. The advantages of using a lower number of elements are to accommodate analytical methods (10) or to use multivariate statistics, such as the Hotelling's  $T^2$  test, without resorting to the requirement of an impracticably large number of replicate measurements (1,3). However, a reduction of the number of variables, whether by deletion or linear combination, results in a loss in the discrimination capability of the overall analytical method. In this study, the discrimination capabilities of several elements are very high, and the concentrations of the ten elements show no significant linear correlations with each other (2). Therefore, a loss of discrimination capability will occur if only some of the elements are used to compare specimens. To assess the magnitude of this loss, t-tests were used to determine the numbers of Type II errors using only the five most discriminating elements: Sr, Fe, Al, Ba, and Ti. These results are shown in Table 3, using the same format as Table 1.

Comparison of the results shown in Tables 1 and 3 at equal overall  $\alpha$  levels indicates that the loss in discrimination capability when reducing the number of variables from ten to five elements is not large. For example, using the equal-variance t-test with an overall  $\alpha$  level of about 0.05, the number of Type II errors increases from two (0.009%) to five (0.02%) when considering the concentrations of only the five most discriminating elements. Similarly, the increase is from 22 to 26 at the overall  $\alpha = 0.01$  level. These results indicate that, despite the fact that each of the ten elements provides good discrimination capability when taken one at a time and that the elements are independent, relatively low levels of Type II errors occur when using only five elements. In part, this low level of Type II errors results because the loss in discrimination by reduction of variables from ten to five is offset by the increase in the individual test  $\alpha$  level needed to attain the same overall Type I error rate. Additionally, these results reflect the fact that glass specimens from different sources almost always display multiple, highly significant differences.

These results might beg the question, why not use only five elements instead of ten? A general answer is that, in cases of comparison of evidentiary specimens, the discrimination capability increases as the number of measured independent parameters increases. Ideally, in forensic comparisons, the number of errors of false association should be minimized at a reasonable rate of false exclusions. Even a small reduction in the number of false associations is desirable, particularly when it is accompanied by minimal analytical effort. By using a greater number of elements in the comparison, the frequencies of both Type I and Type II errors can be decreased. Further, when comparing the concentrations of ten ele-

TABLE 3—Number and percent of indistinguishable pairs using the concentrations of 5 elements (Fe, Al, Sr, Ba, and Ti).

$\alpha$ Level for Each Test	0.01	0.005	0.002	0.001
Indistinguishable Pairs				
<u>Equal Variance</u>				
Number	5	9	26	53
Percent	0.02%	0.04%	0.12%	0.24%
<u>Unequal Variance</u>				
Number	19	52	140	249
Percent	0.09%	0.24%	0.64%	1.15%
Overall $\alpha$ level	0.049	0.025	0.010	0.0050

ments in two specimens, there may be large, readily-recognizable differences in the concentrations of an element that is absent from a five-element comparison protocol. It is particularly important to use a large number of elements when comparing specimens of unknown glass type. Decreasing the number of elements measured from ten to five might be a reasonable consideration when the number of elements is limited by the analytical method used. However, when using ICP-AES or other nearly-simultaneous multielement technique, there is no significant analytical or operational advantage to reducing the number of elements.

The analytical data for three pairs of specimens, which will be used to illustrate several points made in this paper, are shown in Table 4. The first two of the pairs listed are the specimen pairs that are indicated in Table 1 as indistinguishable using the t-test at equal variance and overall  $\alpha = 0.05$ , and also using the Welch modification of the t-test at overall  $\alpha = 0.01$ . For these examples, the equality of variances for each element was tested using an F-test to determine which form of the t-test is appropriate. The third example given in Table 4 represents a pair of specimens that are compositionally similar, but have significantly different refractive indices. The refractive indices of these three specimen pairs are also listed in Table 4. Where RI data are available, they provide a valuable additional point of discrimination. For example, of the twelve pairs of specimens that are indistinguishable using the Welch modification of the t-test at overall  $\alpha = 0.01$ , nine have refractive index data for both specimens. Of these nine pairs, five have refractive indices that differ by more than 0.0002, which, for most glass types, would indicate that they originate from separate sources.

2-K7 and 13-Q1

For all elements, the variances of measured concentrations are equal for the two specimens according to the F test. As shown, all calculated values of  $F$  are less than the critical value of 39. The t-test with equal variance is the appropriate test for these samples, because the variances are not significantly different. However,



TABLE 4—Analytical and summary data for three pairs of specimens that are compositionally similar. The two comparisons that have significantly different variances according to the *F*-test are highlighted in gray.

	RI	Ca	Fe	Al	Mn	Sr	Mg	Ba	Ti	Zr	Na
2-K7	1.5184	6.49	0.095	0.0568	0.00410	0.00504	2.33	0.00210	0.0052	0.0037	10.36
		6.47	0.097	0.0559	0.00378	0.00505	2.34	0.00193	0.0054	0.0037	10.56
		6.35	0.092	0.0583	0.00396	0.00499	2.28	0.00200	0.0060	0.0041	10.42
Mean		6.44	0.095	0.0570	0.00395	0.00503	2.32	0.00201	0.0055	0.0038	10.45
sd		0.08	0.002	0.0012	0.00016	0.00003	0.03	0.00009	0.0004	0.0002	0.10
13-Q1	1.5182	6.23	0.100	0.0686	0.00447	0.00527	2.29	0.00183	0.0068	0.0024	10.54
		6.17	0.097	0.0619	0.00398	0.00524	2.28	0.00192	0.0064	0.0020	10.58
		6.10	0.090	0.0622	0.00414	0.00507	2.24	0.00188	0.0071	0.0037	10.85
Mean		6.17	0.096	0.0642	0.00420	0.00519	2.27	0.00188	0.0068	0.0027	10.66
sd		0.06	0.005	0.0037	0.00025	0.00011	0.02	0.00005	0.0004	0.0009	0.17
F test ( $F_{crit} = 39$ )		1.54	4.22	9.62	2.43	11.26	1.88	3.59	1.24	13.93	2.76
15-K57	1.5192	6.40	0.070	0.0584	0.00114	0.01170	2.41	0.00096	0.0088	0.0085	10.41
		6.26	0.064	0.0525	0.00091	0.01124	2.30	0.00098	0.0085	0.0086	10.30
		6.28	0.066	0.0525	0.00094	0.01136	2.30	0.00099	0.0085	0.0075	10.45
Mean		6.31	0.066	0.0544	0.00100	0.01143	2.33	0.00098	0.0086	0.0082	10.38
sd		0.08	0.003	0.0034	0.00013	0.00024	0.07	0.00002	0.0002	0.0006	0.07
39-K2	1.5190	6.36	0.067	0.0498	0.00130	0.01210	2.28	0.00106	0.0073	0.0063	10.51
		6.31	0.066	0.0455	0.00118	0.01184	2.27	0.00096	0.0079	0.0066	10.57
		6.28	0.066	0.0445	0.00126	0.01205	2.26	0.00100	0.0077	0.0069	10.63
Mean		6.31	0.066	0.0466	0.00125	0.01200	2.27	0.00101	0.0076	0.0066	10.57
sd		0.04	0.000	0.0028	0.00006	0.00014	0.01	0.00005	0.0003	0.0003	0.06
F test ( $F_{crit} = 39$ )		3.73	66.49	1.46	4.19	2.99	64.07	10.86	3.49	3.29	1.42
29b-K1	1.5169	6.02	0.073	0.0371	0.00089	0.00338	2.27	0.00061	0.0063	0.0039	10.18
		6.04	0.072	0.0338	0.00054	0.00330	2.22	0.00057	0.0069	0.0019	10.25
		6.01	0.081	0.0331	0.00094	0.00331	2.17	0.00050	0.0066	0.0022	10.24
Mean		6.02	0.075	0.0347	0.00079	0.00333	2.22	0.00056	0.0066	0.0027	10.22
sd		0.02	0.005	0.0021	0.00022	0.00004	0.05	0.00006	0.0003	0.0011	0.04
106-K1	1.5178	5.65	0.077	0.0455	0.00107	0.00340	2.43	0.00034	0.0054	0.0040	10.18
		5.79	0.081	0.0446	0.00112	0.00345	2.45	0.00038	0.0058	0.0037	10.10
		5.64	0.075	0.0417	0.00121	0.00337	2.43	0.00036	0.0052	0.0042	10.07
Mean		5.69	0.078	0.0439	0.00113	0.00341	2.44	0.00036	0.0055	0.0040	10.12
sd		0.08	0.003	0.0020	0.00007	0.00004	0.01	0.00002	0.0003	0.0003	0.06
F test ( $F_{crit} = 39$ )		25.45	2.42	1.12	9.44	1.16	12.10	7.75	1.47	16.51	2.59

these two specimens are indistinguishable using the *t*-test with equal variance at overall  $\alpha = 0.05$ , but are distinguished by Ca at overall  $\alpha = 0.01$ . The Welch modification incorrectly fails to discriminate among these two specimens, even at the overall  $\alpha = 0.01$  level. The range overlap method clearly differentiates between these two specimens on the basis of Ca, Al, and Ti. The ranges for Sr and Ba also do not overlap for these samples, but the differences are not great enough to be considered significant. The refractive indices of these two specimens are similar, differing by 0.0002.

#### 15-K57 and 39-K2

The variances of the Fe and Mg measurements are unequal for these two specimens, as indicated by *F* values greater than 39. Therefore, the appropriate *t*-test for these samples is equal variance for eight elements and the Welch modification for Fe and Mg. These specimens are not discriminated using the combined equal- and unequal-variance *t*-tests at the overall  $\alpha = 0.05$  level. Using the range overlap test criterion, these two specimens can be differentiated on the basis of Al, Sr, Mg, Ti, Zr, Na, and possibly Mn. In this case, the *t*-test fails to differentiate between these two samples, and

the range overlap test finds seven elements that are different, with one of them, Zr, having substantially different concentrations in the two specimens. The *t*-test is too strict, resulting in an incorrect association for two specimens that are quite dissimilar. As with specimen pair 2-K7 and 13-Q1, the refractive indices of these two specimens are also similar, differing by 0.0002.

#### 29b-K1 and 106-K1

This is an example of a pair of specimens that are differentiated by the equal-variance *t*-test (by Ca, Mg, and Ba) at the overall  $\alpha = 0.05$  level, but are indistinguishable using the Welch modification at the same level of significance. All variances are equal, so the *t*-test is more appropriate than the Welch modification. Range overlap differentiates between the two specimens based on Ca, Al, Mn, Mg, Ba, and Ti. The refractive indices of these specimens differ by 0.0005. For most glass objects, this RI difference is significant and indicates different sources.

As illustrated by the example specimen pairs in Table 4, the range overlap criterion is the most discriminating test of those considered. In this study, use of the range overlap comparison criteria results in

no Type II errors. In fact, most pairs of specimens are differentiated by more than five elements. There is no particularly compelling statistical basis for the range overlap test to be so successful. Unlike other measures of population distribution, such as standard deviation, the range increases with the number of replicates. Therefore, if the number of replicates per specimen were increased, the number of false inclusions would also increase. It is probably a fortuitous circumstance that with three replicates, the range test gives such good error rates, at least as far as Type II errors are concerned. A disadvantage of the range overlap test is the incidence of Type I errors, which is not explicitly known. Statistically, the range of a single set made up of three replicate samples encompasses approximately 1.91 times the standard deviation (11). The average ratio of the range to the standard deviation for all element concentrations in the glass specimens in this study is 1.90, in accord with the theoretical value. Therefore, this range overlap match criterion is tight, resulting in a relatively large Type I error rate. Another disadvantage to the range overlap criterion is that it is more prone to errors caused by outliers than are other hypothesis tests.

## Conclusions

The high degree of discrimination capability obtained when ten element concentrations are compared allows use of the Bonferroni correction to avoid a high frequency of Type I errors on multiple comparisons, while still maintaining a low level of Type II errors. In a prior study of glass compositions, it was stated that lowering Type I errors as needed for the Bonferroni correction will make true differences "almost impossible" to detect (3). The results presented in this paper indicate that this statement is incorrect, at least for glass composition comparisons using ICP-AES data. Specifically, all but 0.009% of specimens from truly separate sources were differentiated using the t-test with a 0.05 overall significance level.

It has also been suggested (1) that, when using a t-test for comparisons of RI in glass, the Welch modification should be used, because the variances generally differ between control and recovered samples. The situation appears to be different for the compositional data included in this study. For most specimen pairs in this study, when the concentrations of an element are approximately equal, the variances of the ICP-AES measurements are not significantly different. It must be noted that the pairs of specimens in this study are not control-recovered pairs, but rather two specimens for which triplicate mg-sized fragments were available. For these samples, it is generally more appropriate to use the t-test with equal variance than the Welch modification for these comparisons. Regardless of the significance level used, the Welch modification always results in a higher number of false inclusions than does the equal-variance t-test. For glass composition comparisons making the assumption of unequal variances and using the Welch modification in all cases is unnecessarily conservative. A better approach would be to use the F-test to assess the equality of variances, and then use the appropriate t-test.

It has also been stated that the range overlap test should not be used, because it results in too many false associations (1). The results of this study indicate that, for glass composition data, this statement is not correct. The range overlap criterion is the most discriminating test of those considered, with no Type II errors and most specimen pairs distinguishable by more than five elements. The range overlap test also has a certain legal appeal and overall simplicity; that is, when the ranges of data points from two separate sources overlap, they cannot be sorted out into two distinct groups. The disadvantage of the range test is not that it results in too many

false associations, but rather that it may result in too many false exclusions, at least when the range is formed by three replicates.

The use of multivariate tests, such as the Hotelling's  $T^2$  statistic for comparison of multivariate data has been recommended (1,3). However, for comparison of ten variables, this test requires a minimum of twelve measurements for each variable. With only three replicates for each specimen in this study, we could not apply Hotelling's  $T^2$  test to the 10-element concentration data. The results of this study indicate that, if the number of measured element concentrations is reduced from ten to five, the reduction in discrimination capability is not large. However, the alternative approach, that of Bonferroni correction to the t-test, seems more appropriate given that one cannot be sure whether elimination of even one element may result in the loss of valuable discriminating information. Another approach that remains to be tested is to reduce the number of variables by linear combination of the ten element concentrations and then use a multivariate test statistic.

The results of this study highlight some of the merits and drawbacks to several methods for interpretation of element concentrations in glass. In cases where triplicate samples can be obtained from both specimens to be compared, as in the data in this study, the most effective comparison method is either the appropriate t-test at an overall 0.05 significance level or the range overlap test. In cases where one or more small, recovered fragments are to be compared to many fragments from a control source, several approaches can be considered. One approach is to characterize the control specimen well by analyzing a relatively large number of representative fragments. The results for each recovered fragment can then be compared individually with this potential source using a Bonferroni-corrected test statistic or a range test. The number of analytical samples appropriate for characterization of the control glass and the specific comparison criteria can be determined using the discrimination capabilities demonstrated in this study and the principles discussed by Sandercock (12). A second approach is to group the recovered fragments based on their RIs and elemental compositions using an appropriate algorithm (13). Comparison of the grouped samples with potential sources can then be made in the same manner as the comparisons in this study. The use of element concentrations to form source-groups among the recovered fragments will be far less prone to grouping errors than is RI data alone because of the wide variations among sources and the good precision within each source.

The evaluation of elemental composition data for evidentiary glass samples conducted in this study 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. The approach presented in this study for the evaluation of matching errors may also be applied to other types of data, such as RI, and other forms of evidentiary materials, besides glass. It has yet to be shown whether other materials may be discriminated to the same degree as that provided by the compositional analysis of glass. The frequencies of Type I and Type II errors for other forms of evidence and analytical parameters depend upon variability across sources or manufacturing production runs, homogeneity of the objects being considered, and measurement precision for the comparison parameters.

## Acknowledgments

The authors would like to thank the members of the FBI Laboratory who performed the analyses and collected the data that formed the basis for this study.

## References

1. Curran JM, Hicks TN, Buckleton JS. Forensic interpretation of glass evidence. Boca Raton, FL: CRC Press LLC, 2000.
2. Koons RD, Buscaglia J. The forensic significance of glass composition and refractive index measurements. *J Forensic Sci* 1999;44:496–503.
3. Curran JM, Triggs CM, Almirall JR, Buckleton JS, Walsh KAJ. The interpretation of elemental composition measurements from forensic glass evidence: I. *Sci Just* 1997;37:241–4.
4. 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.
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. Koons RD, Peters CA, Rebbert PS. Comparison of refractive index, energy dispersive X-ray fluorescence and inductively coupled plasma atomic emission spectrometry for forensic characterization of sheet glass fragments. *J Anal Atom Spectrom* 1991;6:451–6.
7. Miller ET. A rapid method for the comparison of glass fragments. *J Forensic Sci* 1965;10:272–81.
8. Steel RGT, Torrie JH. Introduction to statistics. New York: McGraw-Hill, 1976:145.
9. Mendenhall W. Introduction to probability and statistics, 3rd edition. Belmont, CA: Duxbury Press, 1971:226.
10. Hickman DA. Elemental analysis and the discrimination of sheet glass samples. *Forensic Sci Int* 1983;23:213–23.
11. Massart DL, Vandeginste BGM, Buydens LMC, DeJong S, Lewi PJ, Smeyers-Verbeke J. Handbook of chemometrics and oalimetrics: Part A, data handling in science and technology 20A. Amsterdam: Elsevier, 1997:30–2.
12. Sandercock PML. Sample size considerations for control glass in case-work. *Can Soc Forens Sci J* 2000;33:173–85.
13. Triggs CM, Curran JM, Buckleton JS, Walsh KAJ. The grouping problem in forensic glass analysis: a divisive approach. *Forensic Sci Int* 1997; 85:1–14.

### Additional information and reprint requests:

Robert D. Koons, Research Chemist  
Forensic Science Research Unit  
FBI Academy  
Quantico, VA 22135