November 29, 2017

REFERENCE: Response from Trace/Materials subcommittee to statisticians comments on ASTM E-2927-16

The Materials Trace Subcommittee (MTC) would like to note that the test methods E2330, E2926 and E2927 are highly correlated as all of them address procedures for one of the steps of the forensic examination of glass: the measurement and comparison of the elemental composition of glass fragments. As such, these methods have similar scopes, limitations and implications. The main difference between these test methods is their analytical performance, where E2926 (**u-XRF method approved in the OSAC registry**) is less sensitive and less precise than ICP-MS and LA-ICP-MS (E2330 and E2927) and therefore requires different considerations for the comparison criteria method. Most of the concerns previously expressed by LRC, STG and/or SAC during the two-plus years of the OSAC revision process apply to all these three methods, and therefore some of the concerns that resurfaced during the public comments period of ASTM E2927-16 have been already addressed at different stages during the revision process of the approved E2926-16 method.

Statisticians' Comments on ASTM E2927–16
June 17, 2017

The undersigned members of the FSSB's Statistics Task Group offer these comments to improve the use of the statistical methods and reasoning that are part of the "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 E2927–16).

(1) To avoid misunderstanding on the part of readers who have not heard verbal explanations of the standard and have not read the articles cited, the standard should contain a statement that its choice of the number of replicate measurements and matching rule are not based on statistical theory; hence, the operating characteristics of the rule must be estimated from simulations or empirical tests on same- and different-source glass fragments.

RESPONSE (1) Not persuasive

The sampling and match criteria recommendations described in this test method are derived from empirical tests conducted by several inter-laboratory studies including samples from same -and different-sources. The sampling is also recommended based on a typical fragment size recovered from crime scenes, heterogeneity of the glass, and instrumental variations.

The references added to the revised/reballoted ASTM E2917-16 already addressed this concern. Readers not familiar with the method can refer to the peer review literature.

(2) ASTM E2927–16 has a section devoted to "Interpretation of Results." The standard provides a matching rule for each element in which the mean of three or more measurements on a recovered fragment is compared to an interval of ±4 adjusted standard deviations (ASDs) around the pooled mean of at least nine replicate measurements on at least three fragments from the known sample. (We refer to "adjusted standard deviations" because the value used for the standard deviation (SD) is subject to a floor

of 3%. The literature cited in the standard suggests that this adjustment has a substantial impact on conditional error rates.) The multi-element match rule (in Section 11.1.7) is that "If the mean concentration of one (or more) element(s) in the Recovered fragment falls outside the match interval for the corresponding element in the Known fragments, the element(s) does not 'match' and the glass samples are considered distinguishable."

If the standard is to contain this section on interpretation, the section must include a statement of the conditional error probabilities seen in the existing empirical studies (references 8–11) for the proposed procedure. This statement should refer to the sample sizes for these studies and other major limitations on them. Not only is the estimation of uncertainty or error probabilities essential for every test method, but it is critical in light of the ambitions for the test method delineated in the Introduction to the standard. The Introduction states that

> If the samples are distinguishable in any of these observed and measured properties, it may be concluded that they did not originate from the same source of broken glass. If the samples are indistinguishable in all of these observed and measured properties, the possibility that they originated from the same source of glass may not be eliminated. The use of an elemental analysis method such as laser ablation inductively coupled plasma mass spectrometry yields high discrimination among sources of glass.

The statistical aspects of judgments such as "indistinguishable," "may not be eliminated," and "high discrimination" should be noted and explained in the standard. That is, there should be information for all readers on how high the level of discrimination is and how frequently the test method would classify both nonsources and sources as "indistinguishable."

RESPONSE (2): Previously considered

The terminologies are out of the scope of this standard **test method**. Other standard guidelines are more appropriate documents for defining these terms (e.g., interpretation guideline and the overarching guideline for examination of glass currently being developed within the OSAC MTC). The glass task group within the materials (trace) subcommittee considers that the terms "distinguishable', "indistinguishable" and "discriminating" are well known and well understood. Moreover, these terms appear in the relevant peer-reviewed scientific literature that supports the **test method**, and some are cited within the **test method**. Therefore, adding these terms in the document will not add to the overall substance and utility of the current standard. Moreover, when the **test method** is followed, as written, the measurement outcomes are expected to yield predictable and known analytical figures of merit such as uncertainty boundaries, precision, bias and limits of detection. Finally, the **test method** cites the expected error rates when the recommended match criteria are used for the comparison of glass samples and, importantly, these error rates are based on interlaboratory trials from a substantial number of operational and academic laboratories with expertise in the field of glass analysis and comparisons.

The reasoning supporting our opinion has been provided in previous communications within the OSAC revision process: 1) MTC general comments in response to LRC comments regarding ASTM E2926 and E2330 submitted October 2015 (see response #1 and 2), 2) MTC specific comments in response to LRC comments regarding E2330-12 (see highlighted sections on pages 2, 5,9), 3) MTC specific comments in response to LRC comments regarding E2926-13 (October 2015, see highlighted sections on pages 4 and 9), 4) MTC response to STG recommendations to FSSB on E2330 and E2926, submitted August 2016 (see highlighted sections). The basis for these responses applies to ASTM E2927-16. In addition, the pertinence of adding these terminologies in a **test method** has been discussed by the MTC in several occasions with the STG and LRC during the QIC-coordinated virtual meetings.

(3) The empirical estimation of error probabilities for a test method for making source classifications must rest on a robust series of studies with samples that are representative of case work. ASTM E2927–16 implies that four studies support the ±4 ASD classification rule. Section 11.1 mandates this one procedure (apparently to the exclusion of all others). It states that “The procedure below shall be followed to conduct a forensic glass comparison using the recommended match criteria is [sic] as follows (8-11) ....”

As explained shortly, we are not persuaded that this is a uniquely desirable procedure, and it is clear that not all four of the studies (8-11) recommend its use. If alternatives are scientifically acceptable and not clearly inferior, they should be allowed as well. Therefore, we will briefly discuss each of the four studies. First, Reference (8) is Weis, P., Ducking, M., Watzke, P., Menges, S., and 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, Vol 26, 2011, p. 1273. This article does not recommend the ASTM method, which it calls "n sigma interval around one of the two glasses to be compared." Instead, it recommends the "Modified n sigma criterion with fixed relative standard deviations." Rather than computing a different SD for each known sample, "a fixed relative standard deviation (FRSD) was estimated on the basis of 90 determinations (mean of 3 replicates) of the concentrations in the German glass standard DGG 1 (relative standard deviations from the 90 determinations). In the cases where the relative standard deviation of these 90 determinations was below 3%, the FRSDs were set to 3%."

Even with regard to this "modified criterion," whose operating characteristics may not match the ones of the match rule in the ASTM standard, the results do not unequivocally point to a ±4 FRSD window. Table 8 includes the following findings (rewritten for ease of exposition):

k (in ±*k* FRSD)	False-negative Rate	False-positive Rate
3	133/946 (13.8%)	1/1181 (0.05%)
4	10/946 (1.04%)	2/1181 (0.11%)
5	2 (0.21%)	3/1181 (0.16%).

One might conclude from these numbers that a ±5 FRSD rule is as good as, if not better than, the ±4 FRSD rule.

Furthermore, the extent to which any of these figures provide widely applicable estimates is open to question. The false-negative rates, which drive the selection of ±4, come from tests on a single pane of glass “purchased from a local store in 2005 in Fredericksburg, Virginia, USA ... cut into 144 pieces.” The false-positive rates come from more sources: “62 float glass samples from different countries, manufacturers and production lines, also glass samples from the same production line with different production dates.” Estimating false-positive rates using this set of 62 glass samples, or any set, means that the estimates are based on the characteristics of that particular collection. Applying the overall rate false-positive rate from this ad hoc collection to casework for a particular locale is problematic.

Reference (9) of the standard is 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., and Almirall, J. R., 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, Vol 405, 2013, p. 5393. This study does not supply any conditional error rates at all. It does not recommend the rule in the ASTM standard.

The remaining two studies do recommend the standard's match rule, but, looking at the tables of errors as a function of the number of SDs in the window, it would be reasonable to support a ±5 ASD rule as well. A portion of Table 5 of reference (10) (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., and Almirall, J. R., 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, Vol 28, 2013, p. 1270) can be rewritten (in part) as

	False-negative Rate			False-positive Rate		
k (in ±*k* ASD)	Test2	Test3	Test4	Test2	Test3Test4	
3	0/19	--	56/120 (47%)	0/19	3/126 (2%)	0/60
4	0/19	--	34/120 (28%)	0/19	6/126 (5%)	0/60
5	0/19	--	22/120 (18%)	0/19	14/126 (11%)	0/60

Given these numbers, the false-positive rate seems to driving the choice of the ±4 ASD rule. Moreover, the estimates again are based on limited data. Table 3 indicates that two sources of glass came from one manufacturer for Test 2, five more from this manufacturer for Test 3, and another two from a second manufacturer for Test 4.

Finally, reference (11) is Dorn, H., Ruddle, D.E., Heydon, A., and Burton, B., Discrimination of float glass by LA-ICP-MS: assessment of exclusion criteria using casework samples, Canadian Society of Forensic Science, Vol 48, No. 3, 2015, p. 85. Table 5 of this study includes these results:

k (in ±*k* ASD)	False-negative Rate	False-positive Rate
3	67/2256 (2.97%)	4/6642 (0.06%)
4	6/2256 (0.27%)	7/6642 (0.11%)
5	0/2256 (0%)	12/6642 (0.18%)

The false-negative rates come from pairwise comparisons of 48 fragments from a single “pane of clear, colourless, annealed float glass [that] was purchased from a local glass retail store in 2007.” The false-positive rates come from pairwise comparisons of 82 fragments from “50 samples of architectural glass (33 windows, 17 doors), 24 samples of automotive glass (eight windshields, 14 side/rear windows, two sun roofs) and eight samples of miscellaneous glass (six display cases, two glass tables).”

These rates surely do not preclude the choice of *k* = 5 instead of *k* = 4. Both have small false-positive and false-negative rates. As the authors acknowledge, “for the float glass samples examined in this study, the minimum relative standard deviation approach at ±4 SDmin or ±5 SDmin yields the best results for minimizing both Type I and Type II error rates.” They prefer the former rule only because “the ±5 SDmin criterion produces exclusively Type II errors (false inclusions), whereas the ±4 SDmin criterion shows greater balance between Type I and Type II errors, thereby placing more emphasis on sample discrimination ... .” In other words, the researchers believe that avoiding 7 per 10,000 false positives is more beneficial than incurring 27 per 10,000 false negatives. We do not necessarily disagree, but we must note that it does not seem to be a strongly data-driven conclusion.

Furthermore, even that fine distinction presupposes that this study uses the ASTM matching procedure for the factor that ±*k* multiplies. But SDmin is not the same as ASD. This study does not use the ASTM floor of 3% for adjusting the measured SD for each element. “Here, the RSDmin was set to 4% for Ti, Mn, Rb, Sr, Zr, Ba, La, and Ce, and to 5% for Nd and Pb. The 5% RSDmin for Nd and Pb was necessary given that these two elements generally have a higher actual RSD [relative standard deviation].” Moreover, the number of elements used for the comparison can affect the conditional error rates. This study used 10 trace elements. In Trejos et al. (2013), “Participants using ICP-based methods reported between 10 and 18 element concentrations from the following list: Li, Mg, Al, K, Ca, Fe, Ti, Mn, Rb, Sr, Zr, Sn, Ba, La, Ce, Nd, Hf, and Pb.” Weis et al. (2013) used 18 elements. The last study also used six replicate measurements rather than the nine called for in the ASTM standard. These differences in methodology make it difficult to confidently specify error probabilities for the ASTM rule and to conclude that it is clearly the best method for distinguishing among the possible origins of glass fragments. Therefore, we think the standard should be written so as to identify the ±4 ASD matching procedure as, at most, one potentially reasonable approach for classifying samples according to their possible source.

RESPONSE (3): Previously considered

The reasoning supporting our opinion has been provided in previous communications within the OSAC revision process: 1) MTC response to SAC notification of non-approval justification of ASTM E2927-13 submitted October 2015 (see response #2), 2) The MTC included FSSB recommendations to cite the documents that support the selection of the recommended match criteria in the Standard, the revised document with citations was re-submitted to ASTM ballot and approved in 2016. The MTC also clarified to LRC and STG members during the QIC-coordinated virtual meetings that the method is not meant to be exclusive, and other method that provide similar or better performance in terms of error rates than the recommended method can be used in the practice.

During the revisions and OSAC approval process, it was recommended to the subcommittee to recommend a single match criterion. Nonetheless, the match criterion is a recommendation and does not prevent the use of alternative methods that demonstrate similar or better performance. The use of 4SD was found to provide a balanced low number of false inclusions and false exclusions. Although 5SD produced low error rates, the false inclusion rate was higher in all four studies.

The criteria cited on reference 11 is in agreement with the ASTM 2927 standard method because the section 11.1.3 states “*Calculate a value equal to at least 3% of the mean for each element. This is the Minimum SD*”. The wording “at least” was included to offer flexibility when needed as reported by Dorn et al, where 4-5% was applied for some elements.

This was a match criteria recommended by a world-wide working group, including the authors of the articles who use slight modification to the minimum standard deviation. For instance, the fixed standard deviation reported by Weis et al is a modification of the recommended criteria as a result of their twenty-plus years of historical data that can support the use of fixed values. Currently, there is no other laboratory in the world with such historical data upon which the method relies and, therefore, the method recommended in the standard is more appropriate for most laboratories. Furthermore, interlaboratory studies have demonstrated that using a slight modification to the recommended match interval (as in Dorn and Weis) will not make a large difference in the error rates and conclusions.

(4) The four studies noted in ASTM E2927–16 are all important and valuable contributions to an understanding of how to decide whether glass fragments are “indistinguishable.” We are cognizant of the effort that has gone into the interlaboratory studies, and we do not doubt that the analytical procedures enumerated in ASTM 2927–16 lead to measurements that are precise and informative. But we believe that the standard should be written so as to discourage claims of extremely high discrimination combined with conclusions of “indistinguishable” until more extensive studies of the distribution of elemental

composition in relevant populations are available. At this point, the standard should mention the various approaches that are being used, note that the ±4 ASD and ±5 ASD rules have emerged as superior to certain other procedures in experiments with small numbers of sources of glass. It also should suggest that if that approach is used, conclusions of "indistinguishable" should not be accompanied by estimates of how rare this conclusion would be for all possible sources of glass in the relevant population (because statistically valid estimates are not yet available).

RESPONSE (4): Not persuasive

The comment does not apply to the method because the estimation of how rare an association would be for all possible sources of glass in the relevant population is not discussed in the standard and is outside the scope of the standard.

(5) ASTM E2927–16 is written as if no other methods than the classification rule that it dictates have been proposed. As the Committee knows, however, the use of methods that do not result in match-nonmatch classifications have been described and used in casework (see, e.g., Grzegorz Zadora, Agnieszka Martyna, Daniel Ramos, Colin Aitken Statistical Analysis in Forensic Science: Evidential Value of Multivariate Physicochemical Data, John Wiley & Sons, Chichester, 2014). The failure to mention these alternative approaches gives the standard the appearance of not being fully informed. Therefore, this standard should express a view on such alternatives—for example, that there has not been enough study to warrant their use in this field, or that they are a viable alternative but beyond the scope of this standard. Alternatively, the standard could simply state in the Introduction or in Section 1 that these alternatives exist, but the standard takes no position on their use.

The following ten STG members voted to support the preceding five comments: Banks, David (Duke University), Carriquiry, Alicia (Iowa State University), Guthrie, William (NIST), Iyer, Hariharan (NIST), Kafadar, Karen (University of Virginia), Kaye, David (Pennsylvania State University), Mandal, Abhyuday (University of Georgia), Spiegelman, Clifford (Texas A&M University), Stern, Hal (University of California - Irvine), Zabell, Sandy (Northwestern University). No one abstained or voted against these comments.

RESPONSE (5): Previously considered

The reasoning supporting our opinion has been provided in previous communications within the OSAC revision process: 1) MTC response to SAC notification of non-approval justification of ASTM E2927-13 submitted October 2015 (see response #2). The MTC also clarified to LRC and STG members during the QIC-coordinated virtual meetings that the method is not meant to be exclusive. The test method does not prevent the use of alternative match criteria, as in any other test method, the user can deviate from the recommendation of the method as far as the user can justify and demonstrate the performance and validity of alternative methods are similar or better than the proposed approach.

It is not reasonable to infer that the failure to mention other alternative approaches gives the standard the appearance of not being fully informed. Indeed, the community who wrote this standard test method, and all existing glass standard methods is known for working together towards the progress of the profession. Moreover the studies mentioned by the STG are well known by the glass group, and their authors are currently working with some of the glass experts in the validation and publication of interpretation standards. But, again, it is our position that the interpretation of glass evidence is outside the purpose of this **test method**.

(6) The order in which sample analyses are conducted can affect the rate of matches. Under most circumstances in which drift is present, the recommended order of the measurements will increase the probabilities of both true and false matches. If the recovered fragments and known fragments were to have their positions swapped then perhaps the chance of nonmatches would be increased. Even if the drift is corrected, due to statistical artifacts, the ordering of the sample runs may matter. This should be noted.

The following seven STG members voted to support the preceding comment: Banks, David Banks, William Guthrie, Hariharan Iyer, Karen Kafadar, Abhyuday Mandal, Clifford Spiegelman, and Sandy Zabell. Alicia Carriquiry and David Kaye abstained, and Hal Stern voted against this comment.

RESPONSE (6): Not persuasive:

As indicated in note 2 of the standard, "*A symmetrical arrangement of the analytical sequence of standards and samples is advantageous in minimizing the effects that may result from instrumental drift*". Experimental studies have demonstrated the error rates, both false inclusions and false exclusions, are minimal using the proposed sequence (5,8,11). Furthermore, reference standard glass samples are measured between samples as part of the quality control to monitor any effect of the instrumental drift in accuracy, precision of the measurements and error rates.