Scientific & Technical Review Panel Final Report for OSAC 2022-S-0029 Standard Guide for Interpretation and Reporting in Forensic Comparisons of Trace Materials

Organization of Scientific Area Committees (OSAC) for Forensic Science

STRP Final Report
OSAC 2022-S-0029 Standard Guide for Interpretation and Reporting in Forensic Comparisons of Trace Materials

Organization of Scientific Area Committees (OSAC) for Forensics Science
March 17, 2023

Disclaimer:

Scientific & Technical Review Panel Members

- Chris Hamburg (Chair), ANAB
- Mohammad Almazrouei, Yale-UCL Exchange Scholar
- Madeline Ausdemore, Los Alamos National Laboratory
- Leonora Bender, ATF- Retired
- Chris Bommarito, Forensic Science Consultants
- Jeff Jagmin, Washington State Patrol
- Kenneth Melson, United States Department of Justice – Retired
- Shirly Montero, Arizona State University
- Emily Prokesch, The Office of the Georgia Capital Defender
- Mike Smith, Federal Bureau of Investigation – Laboratory Division

Report Summary:

The Scientific and Technical Review Panel (STRP) for "Standard Guide for Interpretation and Reporting in Forensic Comparisons of Trace Materials" is an independent panel appointed by the National Institute of Standards and Technology (NIST). A STRP is established with a range of experts to consider how well a standard meets the needs of the forensic science, law enforcement, and legal communities, and to recommend improvements to the standards under review. The STRP appreciates the efforts of Cathy Brown, Trace Materials Subcommittee member, while serving as the subcommittee liaison to this STRP during the review process.

The STRP began its review process with a kickoff meeting on April 8, 2022, and concluded with this STRP final report. The panel reviewed the draft standard and prepared comments for the Trace Materials Subcommittee.

Report Components:

The STRP reviewed this draft standard against OSAC's *STRP Instructions for Review* which include the following content areas: scientific and technical merit, human factors, quality assurance, scope and purpose, terminology, method description and reporting results. The details below contain a brief description of each reviewed content area and the STRP's assessment of how that content was addressed in the draft OSAC Proposed Standard.

1. **Scientific and Technical Merit:** OSAC-approved standards must have strong scientific foundations so that the methods practitioners employ are scientifically valid, and the resulting claims are trustworthy. In addition, standards for methods or interpretation of results must include the expression and communication of the uncertainties in measurements or other results.

1.1 **Consensus View** - The draft standard proposes a qualitative interpretation scale to use in the reporting of trace evidence examinations. The STRP notes that the standard clearly states that the reporting of opinions is to meet or exceed the requirements of ASTM E620 *Standard Practice for Reporting Opinions of Scientific or Technical Experts*, and the case file should include sufficient detail so that another examiner with expertise in the applicable evidence type would come to the same opinion.

Minority View - This guide should not be used unless and until the methods are validated, which must be demonstrated through multiple appropriately designed studies that empirically establish the accuracy, repeatability, and reproducibility of the method for the intended use under similar circumstances as it will be used. Validated methods should then only be used for the factor space that it has been tested and validated for.

1.2 **Consensus View** - The proposed interpretation scale is to facilitate the assessment of the significance of the results obtained with validated methodology used for the comparison of the items. Each evidence-specific section gives a short explanation of the nature of the examination. When possible, the draft also cites thoroughly the standard test methods, practices, and guides for the analysis of fibers, glass, hair, paint, and tape. In addition, it provides clear examples regarding how the typical analytical scheme used during the examination of such a type of evidence could influence the interpretation and the degree of association. Examples of how other factors could weight on the associations are also shown in each of the evidence specific sections; when an association can be made stronger (e.g., a cross transfer of

paint evidence) or when an association is weaker (e.g., association of two denim type fibers).

Minority View #1 - This guide should not be used unless and until the methods are validated, which must be demonstrated through multiple appropriately designed studies that empirically establish the accuracy, repeatability, and reproducibility of the method for the intended use under similar circumstances as it will be used. Validated methods should then only be used for the factor space that it has been tested and validated for.

Minority View #2 - When possible, the draft cites the standard test methods, practices, and guides for the analysis of fibers, glass, hair, paint, and tape. However, the STRP notes that the proposed standard heavily relies on a single 2020 review paper that contains more than 450 references. This might make the draft standard unbalanced in terms of citing the wider, relevant literature. In addition, citing a review paper on many occasions throughout the standard might make it impractical for the end users to look up the specific papers within 450+ references. Hence, it is suggested to include other papers for balance, and when possible, to include references that address the specific points in the standard.

1.3 **No Consensus**

View #1 (6/10) - The STRP observes that although the cited methods are the current accepted and validated methods for the examinations, the use of the interpretation scale itself has not been validated. The STRP considers this draft standard to be a useful guide to qualitatively communicate the significance of an association in trace evidence analysis, but strongly recommends its validation.

View #2 (3/10) - Similar interpretation scales have been in use in trace evidence for the past fifteen years. The use of an interpretation scale provides a systematic and standardized protocol to assess and relay significance in comparative analyses. Factors stated in the guide that increase or decrease significance are well grounded in the scientific literature. Some panelists are uncertain what specifically would be validated, other than assessing whether different analysts would come to the same association level when assessing the same evidence. However, an inter laboratory study was conducted by Mehltretter and Trejos (2021) showing a high level of agreement regarding the significance of results in comparative examinations when using the proposed guide among practitioners who were given scenarios for paint examination.

Meltretter, A, Trejos, T. (2021) Lessons Learned from Paint and Tape Interlaboratory Exercises: Do systematic approaches lead to examiner consistency? Presented at the Forensic Online Symposium: Current Trends in forensic Trace Analysis. (Note:

View #3 (1/10) - This guide should not be used unless and until the methods are validated, which must be demonstrated through multiple appropriately designed studies that empirically establish the accuracy, repeatability, and reproducibility of the method for the intended use under similar circumstances as it will be used. Validated methods should then only be used for the factor space that it has been tested and validated for.

2. **Human Factors:** All forensic science methods rely on human performance in acquiring, examining, reporting, and testifying to the results. In the examination phase, some standards rely heavily on human judgment, whereas others rely more on properly maintained and calibrated instruments and statistical analysis of data.

2.1. **Consensus View** - This draft proposes language and an interpretation scale to describe the significance of results in trace evidence analysis. The scale contains five categories with three subcategories and attempts to clearly describe each in a qualitative manner. The draft also provides specific examples on the use of the scale when reporting results of fiber, glass, hair, paint, and tape examinations. The STRP is positive about the proposal.

Minority View - One of the panelists agrees with the critiques but disagrees with the possibility of publishing the standard with the intention to improve upon it and address the shortcomings later. This document, once published, will be relied on, without any caveats about future improvements and the improvements that are necessary, not aspirational. There is no process to ensure that improvements are made later; the time to make those improvements is now, and therefore, should be made prior to publication or use of this standard. It is unclear whether 2.3 is accepting or endorsing publication. The panelist wants stronger language about the scale than the comment has but would vote to accept 2.1-2.3 so long as it is not accepting the use of the standard prior to improvements.

2.2. **Consensus View** - The lack of quantitative indicators of the category boundaries makes the decision-making process vulnerable. The STRP believes that this draft partially addresses issues related to human factors and performance. Since the processes of ‘interpretation’ and ‘reporting’ rely heavily on forensic practitioners’ decision-making, human factors issues are expected, especially in difficult decisions such as the differentiation between *association with highly discriminating characteristics* and *association with discriminating characteristics*. Despite the attempts in the standard to offer examples of each type of association category, there is a lack of objective and comprehensive criteria for evaluation. As such, these decisions rely on the knowledge and experience of the practitioners themselves and are prone to cognitive bias and both inter- and intra-examiner variability. To be clear, criteria for assessing the other association categories are also lacking. Of mention is

the category of inconclusive decisions where the possibility of unjustifiably reporting these decisions may have practical implications (see Dror & Langenburg, 2019, for details).

Minority View #1 - This guide provides specific examples and background references that do provide "objective and comprehensive criteria for evaluation". While a quantitative assessment of significance would be preferable, the underlying data to perform a quantitative assessment in trace evidence areas, with the exception of glass, is not feasible due to an ever-changing population of materials and variation of populations both geographically and temporally. While a "quantitative" Bayesian assessment would on the surface seem to be more objective, the underlying statistics used in such assessments can be numerical estimates based on the same factors that increase or decrease significance.

Minority View #2 - While there is a quantitative aspect lacking in the standard, it may not be possible to establish quantitative thresholds for many types of trace evidence. Providing quantitative thresholds does not necessarily eliminate the influence of cognitive bias or inter-/intra-examiner variability.

2.3. **No Consensus**

View #1 (6/10) - It might be the case that this standard needs to be published as-is so that it reaches out to the users in a timely manner, and improvements to the document can be made at a later stage. If this is the case, it is recommended that explicit statements are added to outline the limitations of the standards and the potential for cognitive bias. In addition, the STRP recommends including a statement with the need to consider —e.g., as part of the quality manual— the optimization of performance in the decision-making process by considering factors of fatigue, tiredness, low stress, and boredom (see Almazrouei, M. A., Dror, I. E., & Morgan, R. M., 2020, for details).

- Dror, I. E., & Langenburg, G. (2019). "Cannot decide": The fine line between appropriate inconclusive determinations versus unjustifiably deciding not to decide. Journal of Forensic Sciences, 63, 1–6. https://doi.org/10.1111/1556-4029.13854

- Almazrouei, M. A., Dror, I. E., & Morgan, R. M. (2020). Organizational and human factors affecting forensic decision‐making: Workplace stress and feedback.

View #2 (3/10) - Human factors assessments are beyond the scope of this standard. The panelists do not believe it is appropriate to include a decision-making process based on these factors (e.g., fatigue, tiredness, low stress, boredom) in this standard.

View #3 (1/10) – See Minority View for 2.1.

3. **Quality Assurance:** Quality assurance covers a broad range of topics. For example, a method must include quality assurance procedures to ensure that sufficiently similar results

will be obtained when the methodology is properly followed by different users in different facilities.

3.1. **Consensus View** - This draft standard creates a standardized set of reporting language to use for the myriad types of trace evidence that might be encountered in forensic casework. It also suggests where to use it in the interpretation process, and which other information must be provided to support the reported significance of the results using the proposed interpretation scale.

The draft also emphasizes the need to include the interpretation scale in the report in its entirety. The STRP agrees with this requirement and recommends making inseparable a comment on the maximum significance possible for the method used with respect to the maximum significance in the field. This would contextualize the capability of the reporting laboratory with respect to a consensus interpretation scale.

The STRP recognizes that the uniform language proposed represents a leap forward for the trace evidence discipline and its recommendations are intended for use on subsequent revisions and are not intended to oppose publication on the OSAC Registry at this time.

Minority View – One panelist agrees that studies should be conducted, but to the extent the consensus view suggests or endorses the use of standards and methods prior to these studies being conducted, the panelist strongly disagrees. Appropriately designed studies, designed and implemented by independent, qualified research scientists must be conducted, and empirical support must be established prior to the implementation of the proposed standard or the methods and scales described therein. It is a necessary predicate.

3.2. **No Consensus View**

View #1 (5/10) - Although the STRP supports the proposed uniformity when using the interpretation scale, this uniform language does not have a robust foundation of research to support its use. The STRP recommends two quality assurance-based studies. The STRP recommends studies to ensure that the end users of the report (e.g., juries, judges, investigators, lawyers) understand the language used and that the intended information is being conveyed. The STRP recommends studies to ensure that there is limited inter- and intra-author variability in the “level of association” selected based on the data from a particular examination.

View #2 (4/10) - Similar interpretation scales have been in use for over fifteen years.

The publication of the entire scale with associated definitions in forensic reports provides the reader context to different levels of significance that can be reached in the assessment and the reasoning for any increased or decreased levels of significance. The conveyance of this significance assessment, along with the context in the report, is an important step forward as such information is often conveyed in testimony and not the report itself. Additional studies are not required as the

foundation for assessment is clearly conveyed in the guide. While there may be limited instances of "inter and intra-author variability" in significance assessment, this variation should be limited by the comprehensive examples in the guide and some differences may be appropriate given differences in the knowledge and experience of the examiner. Please refer to the study sited in section 1.3.

View #3 (1/10) – One panelist agrees that studies should be conducted, but to the extent the consensus view suggests or endorses the use of standards and methods prior to these studies being conducted, they strongly disagree. Appropriately designed studies, designed and implemented by independent, qualified researcher scientists must be conducted and empirical support must be established prior to the implementation of the standard or the methods and scales described therein. It is a necessary predicate.

4. **Scope and Purpose:** Standards should have a short statement of their scope and purpose. They should list the topics that they address and the related topics that they do not address. Requirements, recommendations, or statements of what is permitted or prohibited do not belong in this section.

4.1. **Consensus View** - After analyzing the Scope section of the proposed standard, the STRP agrees that the proposed language contained in the document substantially describes the scope and purpose of the standard. The Scope discusses the topics addressed and the specific sub-disciplines that are highlighted. It also describes the purpose for which the proposed standard has been developed.

4.2. **Consensus View** - The statement of purpose leaves testimony as an unaddressed topic**.** The purpose of the proposed standard does not include the use of the standard by other forensic science stakeholders such as lawyers, judges, and researchers.

Minority View - The interpretation scale consists of a series of reasonable statements; therefore, studies to determine whether the end-user is able to understand the language in the interpretations scale and the intended information are not necessary.

4.3. **No Consensus View**

View #1 (6/10) - When the proposed standard states that it "can be applied to other trace materials", it provides no caveat as to the appropriateness of other trace materials not highlighted in the standard, the identity of those other trace materials, or

examples of what trace materials would not be susceptible for use of this standard.

View #2 (4/10) - Since trace evidence is a 'catch-all' discipline for manufactured materials, there are numerous materials that may be analyzed but not expressly covered in the guide. Examples of these include, but are not limited to, adhesives, glues, plastics and other polymeric materials, greases, lubricants and waxes. Comparison of questioned and known items of these man-made, mass-produced

materials will usually be an association (unless a physical match or elimination is the result of analysis). The guide provides a general framework for the assessment of these materials not explicitly covered in the guide.

4.4. **Consensus View** - The proposed standard states it is intended for use by practitioners who have "discipline-specific training" and provides examples of ASTM standards and work items for paint, tape, and hair, but not fibers (WK78748), and it does not mention there is no discipline-specific training developed for glass at this time.

Minority View - While section 1.4 in the draft standard lists several ASTM practices as examples, some panelists do not believe the list is intended to be inclusive of all discipline-specific training, which would be outside the scope of the document. The omission of ASTM WK79747 (*Standard Guide for Forensic Examination of Fibers*) and ASTM WK72932 (*Standard Guide for Forensic Glass Analysis and Comparison*) does not affect the validity of section 1.4 which states, "*[t]his standard is intended for use by competent forensic science practitioners with the requisite formal education, discipline-specific training and demonstrated proficiency to perform forensic casework*". The examples listed in the section could be amended with "Including but not limited to" language if the intent of inclusion of the listed practices as examples is not clear or all the listed references can be removed.

4.5. **Consensus View** - The proposed standard states it is intended for use by practitioners who have "demonstrated proficiency to perform forensic casework." The quoted language does not describe the parameters of "demonstrated proficiency" in temporal terms if the language pertains to proficiency testing. The test to "perform forensic casework" normally taken by a practitioner is a "competency" test.

Minority View - "Competency tests" are given after an examiner has completed his/her training to evaluate if they are competent to proceed with casework. "Proficiency tests" are given periodically to the examiner to ensure that they are still performing their work competently. If the laboratory is accredited, the frequency of the proficiency tests are dictated by the accrediting body; however the laboratory may administer tests more frequently than the accrediting body requires (but not less). Therefore, the frequency of the tests given to determine examiner competency is outside the scope of this document.

5. **Terminology:** Standards should define terms that have specialized meanings. Only rarely should they give a highly restricted or specialized meaning to a term in common use among

the general public.

5.1. **Consensus View** - The STRP finds that the draft standard defines appropriate terms with specialized meaning utilized within the standard and appropriately references ASTM E1732:22 for terms that are commonly employed for general forensic examinations, to avoid conflicting terminology between ASTM documents.

5.2. **Minority View** - None

6. **Method Description:** There is no rule as to the necessary level of detail in the description of the method. Some parts of the method may be performed in alternative ways without affecting the quality and consistency of the results. Standards should focus on standardizing steps that must be performed consistently across organizations to ensure equivalent results. Alternatively, standards can define specific performance criteria that are required to be demonstrated and met rather than specifying the exact way a task must be done. For example, it may be enough to specify the lower limit for detecting a substance without specifying the equipment or method for achieving this limit of detection.

6.1. **Consensus View** - The STRP considers that the proposed standard meets the Method Description requirement. While this standard is not about how to analyze a sample of trace evidence, but rather how to interpret and report the findings associated with trace material comparisons, this standard provides ample and sufficient detail to ensure that analyses are consistently and appropriately reported.

Minority View #1 - One of the panelists agrees with the consensus view but believes validation of methods should be a requirement prior to publication of the proposed standard.

Minority View #2 - The Method Description requirement focuses on the consistency of the results. This standard guide is using examples to define the boundaries among the different levels of the interpretation scale. With this choice, the provided examples should be exhaustive “to ensure equivalent results” in ALL cases. The alternative would be to select a performance criterion (e.g., a variation of no more than one association level for multiple analysts), and that is only possible after validating the use of the interpretation scale. Neither of the two alternatives are provided by the standard guide, therefore this standard does not meet the Method Description requirement.

6.2. **Consensus View** -This standard not only provides a process for interpreting findings from trace material comparisons and for clearly communicating the significance of these findings, but also includes definitions of the various levels of associations an examiner could reach. In addition, by including detailed examples for various types of trace evidence at each association level, the standard minimizes ambiguity by demonstrating how these levels can be used in the different trace disciplines.

Minority View - See comment 6.1 for Minority View #1.

7. **Reporting Results:** Methods must not only be well described, scientifically sound, and comprehensive, but also lead to reported results that are within the scope of the standard, appropriately caveated, and not overreaching.

    7.1. **Consensus View** - The STRP notes that the standard is primarily about reporting the results of analyses and the use of the Interpretation Scale. The standard includes many examples, across common trace evidence disciplines, of variables that may lead to

higher or lower significance levels in interpretation. In addition, these examples provide thorough guidance so that different examiners using the standard should reach the same association level. Also included in the standard are helpful examples of wording that may be used when reporting the results of analysis.

Minority View #1 - One panelist believes that there are many examples provided, but is not certain that they are enough to eliminate ambiguity.

Minority View #2 - One panelist agrees with the requirements regarding what needs to be documented and maintained in the case file, but does not agree with using a scale to report results unless the scale has been validated and there is empirical support for the categories and use.