Scientific & Technical Review Panel Final Report for 2022-S-0001 Standard Guide for Image Comparison Conclusions/Opinions

Organization of Scientific Area Committees (OSAC) for Forensic Science

STRP Final Report
2022-S-0001
Standard Guide for Image Comparison Conclusions/Opinions

Organization of Scientific Area Committees (OSAC) for Forensics Science
July 14, 2022

Disclaimer:

This report was produced by an independent Scientific and Technical Review Panel (STRP). The views expressed in the report do not necessarily reflect the views or policies of the U.S. Government. Visit the OSAC website for more information on OSAC's STRP process.

Scientific & Technical Review Panel Members

- Andrew Cohen, University of Massachusetts Amherst
- Brandon Epstein, Middlesex County Prosecutor's Office
- Cami Fuglsby, South Dakota State University
- Melissa Gische, FBI – Laboratory Services
- Steve Johnson, Ideal Innovations
- Julia Leighton, Public Defender (Retired)
- Krista Rembold, FBI - Biometric Services Section

Report Summary:

The STRP has reviewed and discussed this draft standard but only reached consensus on the Method Description content area section. As a result, the other sections have been updated to include the two different views represented by the STRP and the number of votes each view received.

The Scientific and Technical Review Panel (STRP) for "Standard Guide for Image Comparison Conclusions/Opinions" 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 Lora Sims, Digital/Multimedia Scientific Area Committee chair, while serving as the subcommittee liaison to this STRP during the review process.

The STRP began its review process with a kickoff meeting on December 3, 2021, and concluded with this STRP final report. The panel reviewed the draft standard and prepared comments for the Facial Identification 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 **View 1 (4 votes)** – The standard outlines the opinion categories that shall be reached when conducting image comparisons but does not address the processes used to reach those conclusions, which will be covered in separate standards. The STRP acknowledges that image comparisons cover multiple disciplines, some with more research than others. When opinions are provided, the standard requires that reference be made to supporting empirical studies for those disciplines with existing research. For those disciplines without empirical studies, the standard requires that opinions offered in those areas note the absence of research.

> *4.2 Opinions shall include reference to any empirical studies or note the absence of studies for a given type of evidence and interpretation.*

While the standard includes references to existing black box, repeatability, and reproducibility studies, it also notes that there is limited research in these areas. The standard further explains that the requirements are based on the information available at the time of publishing.

> *1.1.4 This standard is based upon practical experience, research, and resources available at the time of publishing. Published research demonstrates that trained practitioners are effective in image comparison. At present, there is limited research is that addresses the ability of practitioners to reproducibly apply the opinion categories listed in this standard or the ability of laypersons to interpret their meaning.*

Finally, the standard includes limitations associated with the opinions and prohibits use of problematic terminology, such as 'individualize.'

> *5.5.1 A practitioner shall confine an opinion to the support-based categories provided in this guide and shall not opine that two items (e.g., faces, vehicles, clothing, skin details) originate from the same source to the exclusion of all others. A practitioner shall not use terms in the stated opinion such as 'individualize' or 'individualization.' A practitioner shall not express an opinion as an absolute certainty.*

In doing so, the standard appropriately addresses limitations for those disciplines with limited research and requires transparency in relating that information to stakeholders.

1.2 **View 2 (3 votes)** – With respect to scientific and technical merit our comments focus on the validity of the proposed scale. First, whether the scale has been empirically tested. Second, is the scale internally consistent in the analysis it is suggesting to the

examiner. In sum, we find that the scale has not been validated and inappropriately combines the three main interpretation methods used in forensic science, and thus fails to provide practitioners with adequate guidance for consistent application.

Validity of the scale

This document sets forth a 5-point scale for image (e.g., people, objects, scenes) comparisons conclusions making the following statement about the merit of the scale.

> *"This standard is based upon practical experience, research, and resources available at the time of publishing. Published research demonstrates that trained practitioners are effective in image comparison. At present, there is limited research that addresses the ability of practitioners to reproducibly apply the opinion categories listed in this standard or the ability of laypersons to interpret their meaning."*

In support of this statement, three studies are cited. One study evaluated the "detailed, behavioral properties of face matching performance in two specialist groups: forensic facial examiners and super-recognizers" using a 7-point identity judgment scale. Another study compared the performance of a small group of trained examiners, other forensic employees (managers etc.) and college students using a 5-point verbal scale that differs from the one suggested here and did not test for reproducibility and repeatability. A third study tested 6 FBI examiners and 5 FBI interns on their respective abilities to identify the specific make, model and year range of vehicles.

This body of research provides, at most, limited support for the statement that trained practitioners outperform lay persons in image identification. The research provides no data on the accuracy, repeatability, or reproducibility of results when practitioners apply the proposed scale.

The document implicitly acknowledges the absence of data on repeatability and reproducibility and the absence of any research on implementing the 5-point scale when it states that "[o]rganizations should ensure appropriate procedures are in place to promote consistent implementation of their opinion scales" but provides no guidance on how this should be done.

The document also acknowledges the importance of validating scales when it states that "the opinion categories 'Support for Exclusion' and 'Support for Common Source' may be subdivided into more specific levels of relative support when empirical research demonstrates that examiners can accurately and reliably gauge the more finely grained categories." We agree further refinement would require empirical

research, just as we respectfully suggest that prior to being placed on the OSAC Registry the proposed scale (or any other scale) needs the same.

Creating a conclusion scale to encourage standardization and allow for future research is desirable. That the proposed scale prohibits individualization or source determination is to be commended as an advancement in the understanding of uncertainty inherent in any method. Incorporating the scale into an ASTM standard, combined with adding the validation of it as a subcommittee research need, would be significant steps towards developing the needed empirical research to validate a conclusions scale for the field.

If the FSSB determines this standard should be placed on the Registry, the standard should include an unequivocal statement that the scale has not been validated (and that no scale has in this field). Because the OSAC Registry is advertised as "a repository of high-quality, technically sound published and proposed standards for forensic science", failure to clearly state that the scale has not been validated while placing the standard on the Registry would be misleading.

But, of even greater concern, is that this 5-points scale confuses the presentation of alternative hypotheses and haphazardly uses language associated with a likelihood ratio (LR) approach, a Bayesian approach and a binary approach when defining the various conclusions.

The presentation of hypotheses

As explained below, the document's use of "exclusion" (See section 4.3 and the descriptions in Section 5) on occasion confuses "hypothesis" or "proposition" with a decision.

To better understand this comment, consider the basic formulation of the Scientific Method. One formulates a hypothesis and sets up an experiment to test the hypothesis. If that hypothesis is found not to be true, then there must be another explanation.

Suppose you are comparing two pictures of red cars, where in picture #2 it is known who owns the car. Hypotheses are:

> Hypothesis 1: The red car in picture #1 is the same red car in picture #2.
>
> Hypothesis 2: The red car in picture #1 is a different red car than the red car in picture #2.

It is important to be specific about what you are looking at, and what the alternative hypothesis is (Hypothesis 2).

Section 4.3 says the hypotheses are either "same source" or "exclusion". Exclusion is not an "alternate hypothesis" as it claims here. Exclusion is a decision reached. Going back to the Scientific Method example, the alternate hypothesis must be another explanation for the photos. In our example, the alternative explanation is that there is a different red car. It is important to convey this to the trier-of-fact as saying there is another explanation out there, or there is another car they are looking for.

Hypotheses should always be stated "positively". "Negative" words like "exclude" or "these do not share the same source", should not be used. If negative words are used, then the focus remains on the given source and evidence and not on another possible source/explanation.

Exclude is not an explanation, exclude is a decision. In the dictionary, the word exclude is a verb, meaning an action taken. Here the action is the decision on the hypothesis that they share a source.

Note that in Section 5, each of the descriptors follow the same pattern, only changing the order of the hypotheses or changing the descriptor of the "probable".

> *"5.1 Exclusion: …The observed characteristics are much more probable given the proposition that the images depict a different source, rather than the proposition that they depict the same source."*

The hypotheses listed here are 1: The images depict a different source, and 2: [the images] depict the same source. These hypotheses do not line up with what was listed earlier in Section 4.3, either "same source" or "exclusion". Also, the title of this opinion is "Exclusion", and so having an opinion and hypothesis that are the exact same is causing dissonance.

The fix to this point is straight forward. Change section 4.3 to say "…of same source and different source" and then use this language throughout the document, including changing the title of "Exclusion" to "Strong Support for Different Source". But this is not the only problem with wording in Section 5.

<u>LRs, Bayesian or a two-stage approach</u>

There are three main interpretation methods to use in forensic science, but in this instance the Opinions section combines the three haphazardly. The titles suggest a Bayesian approach. The occasional use of exclusion suggests a two-stage approach and the second sentence of each opinion suggests an LR comparing two alternative

hypotheses.

As discussed above, Section 4.3 as currently drafted sets forth a Classical/Kirkian/Two-Stage Approach (similar to a p-value). A decision-making approach that, based on conversations with practitioners, appears to be a dominant approach in many jurisdictions. And the first sentence of, for example Inconclusive, "an opinion that there is insufficient information to distinguish between a common source and an exclusion" continues with this two-stage approach.

On the other hand, the second sentence of each definition in Section 5 provides an approach in which two hypotheses (same source, different source) are considered.

> *"5.1 Exclusion: …The observed characteristics are much more probable given the proposition that the images depict a different source, rather than the proposition that they depict the same source."*

But the titles of each Subsection in Section 5.1 (other than Exclusion which as discussed above is a two-stage approach) are the same as the ENFI scale when a Bayesian approach is used. The Bayesian posterior probabilities will read along the

lines of "The [posterior] probability that the hypothesis is true, given the evidence". This approach is reflected in the titles used in Section 5. For example, "Support for common source" reads that there is support for the hypothesis that the cars in the two images share a common source.

A Bayesian probability follows a specific formula that involves relative likelihoods (hence the descriptors reading like likelihoods), but their result only addresses one hypothesis, common source. It is called a Bayesian posterior probability because it uses Bayes theorem, which involves the relative likelihoods. Similar to the reason why you cannot exclude when using LRs, you also cannot exclude using relative likelihoods. Thus, exclusion is also an inappropriate category for a Bayesian approach.

Any of the three approaches could be employed, but they must be presented separately and with clear definitions and guidance on their application and expression that is internally consistent with the approach chosen.

Recommendation: In close collaboration with statisticians (for accurate definition and expression of statistical approaches), human factors resources (for expertise on what guidance to provide to promote consistent application etc.) and the Legal Task Group (for expertise on the use of probabilities in the legal arena etc.), a decision should be reached about whether to use relative likelihood opinions, Bayesian opinions, or exclusionary/Kirkean/two-stage approach opinions. One could include more than one

approach, but they should be in different sections and not intermingled. ENFSI standards have examples for both a likelihood-based scale and a Bayesian scale. Paul Kirk's book Crime Investigation includes descriptions of the exclusionary approach (i.e., Kirkean).

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. **View 1 (4 votes)** – The STRP observed that the standard recognizes the subjectivity involved in image comparisons.

    > *4.5 Image comparison is a subjective practice in nature. Organizations should ensure appropriate procedures are in place to promote consistent implementation of their opinion scales.*

    The STRP further notes that many human factors elements should be addressed in a test method standard outlining how to perform an examination and may be less applicable in a standard that only describes the opinions that can be reached. To that

end, the STRP recommends that the scope be modified to state that human factors are not covered in this standard.

2.2. **View 2 (3 votes)** – For the reasons detailed above, the Opinions section of this Standard Guide does not provide clear and actionable guidance on what procedures promote consistency or how to develop procedures to promote consistency. Procedures for consistent application is a pillar of reliability that cannot be ignored or delayed. A scale that produces results that are not consistently repeated or reproduced cannot be accurate.

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. **View 1 (4 votes)** – The STRP notes that quality assurance should be addressed in a test method standard outlining how to perform an examination and may be less applicable in a standard that only describes the opinions that can be reached. As a result, the STRP recommends that the scope be modified to state that quality assurance is not covered in this standard.

    3.2. **View 2 (3 votes)** – For the reasons detailed above, the Opinions section of this Standard Guide does not provide clear and actionable guidance on what procedures

promote consistency or how to develop procedures to promote consistency. And as noted above, consistent results are a pillar of reliability.

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. **View 1 (4 votes)** – As documented under Human Factors and Quality Assurance, the STRP recommends that the scope be modified to state that neither human factors nor quality assurance is addressed in this standard. The STRP observed that the scope clearly states that the standard does not cover documentation or reporting of the opinions. It also clearly states the intent is to set requirements for the opinions that can be reached but does not describe the process used to reach those opinions.

1.1 This standard defines conclusion (hereafter "opinion") categories that shall be reached by a practitioner performing comparisons of people, objects, or scenes captured in images (e.g., faces, vehicles, clothing, skin details), regardless of the process by which opinions are reached (i.e., the examination methodology).

1.1.2 This standard does not address the documentation or reporting of an opinion

The STRP did note "shall" statements in the Scope but was informed that ASTM, the intended SDO for this standard, allows for "shall" statements in the scope.

4.2. **View 2 (3 votes)** – For reasons discussed above, the Scope should clearly state that the scale proposed has not been validated. It could also include a justification for moving toward a more uniform approach that would allow for empirical research to assess the accuracy, repeatability and reproducibility of results using the scale. The Scope should also state what approach(s) is being employed (e.g., LRs, two-stage).

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. **View 1 (4 votes)** – The STRP acknowledges the standard's attempt to transition the community from using the term "conclusion" to using the OSAC Preferred Term "opinion."

The STRP found the terms in the terminology section of the standard to be acceptable.

The STRP appreciates the standard's inclusion of prohibited language. The STRP discussed whether to recommend adding the term "match" to section 5.5.1. The SC liaison explained that "match" was not included because it is still used by some database vendors. Some STRP members still recommend inclusion of the term to encourage change, while others feel this is unnecessary as the intent of the section is still being met. Similarly, some STRP members recommend that "came from the same source" should be explicitly prohibited as a separate statement from "came from the same source to the exclusion of all others."

> *5.5.1 A practitioner shall confine an opinion to the support-based categories provided in this guide and shall not opine that two items (e.g., faces, vehicles, clothing, skin details) originate from the same source to the exclusion of all others. A practitioner shall not use terms in the stated opinion such as 'individualize' or 'individualization.' A practitioner shall not express an opinion as an absolute certainty.*

5.2. **View 2 (3 votes)** – See discussion under Scientific and Technical Merit about using a consistent approach to the conclusions in the scale.

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 (5 votes)** – The STRP determined that this topic was not applicable to this draft standard because the standard outlines the opinion categories that shall be reached when conducting image comparisons but does not address the processes used to reach those conclusions, which will be covered in separate standards.

6.2. **Minority View (2 votes)** – No test method is addressed in this document. With respect to the proposed approach(s) to interpreting or expressing results, see the discussion under Scientific and Technical Merit.

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. **View 1 (4 votes)** – The standard outlines the opinion categories that shall be reached when conducting image comparisons but does not address how to report those

opinions, which will be covered in a separate standard.

> *1.1.2 This standard does not address the documentation or reporting of an opinion (FISWG Minimum Guidelines for Facial Image Comparison Documentation, SWGDE Technical Overview for Forensic Image Comparison).*

For discussion on the results (i.e., opinions) themselves, see Scientific and Technical Merit section above.

7.2. **View 2 (3 votes)** – The document does not provide practitioners a coherent definition for the opinions permitted, thus creating future difficulties for developing a coherent approach for reporting those opinions. See comments above under Scientific and Technical Merit.