Guidance for OSAC Subcommittees Drafting and Updating Opinion Scale Standards for Comparative Examinations

1. Introduction

In 2024, the Forensic Science Standards Board (FSSB) formed a task group on opinion scales to assist the board in developing guidance for subcommittees that are developing standards for reporting and testifying to inferences from observations/data acquired in comparative examinations. In comparative examinations, two or more physical or digital items are examined to investigate which, if any, of the items might have originated from the same source. This guidance document contains requirements and recommendations for developing strength-of-evidence opinion-scale standards for source-level opinions that appear in reports or testimony intended to communicate the results of comparative examinations in criminal or civil proceedings. The document defines relevant terms in Section 2. It presents requirements and recommendations in Sections 3-9. Although it requires strength-of-evidence scales over truth-of-hypotheses scales, the appendix gives examples of both types of scales to aid standards writers in distinguishing between these types of scales.

This document is aligned with ISO 21043, Forensic sciences - Part 4: Interpretation, published in June 2025. In Part 4, Section 5.1, ISO identifies the various types of questions asked in forensic science, including the "question of source." In Section 5.2, ISO articulates the types of interpretations that can be performed, including investigative interpretations or evaluative interpretations. In Section 6, ISO explains that these interpretations lead to various types of opinions, which include likelihood ratios, categorical opinions (see definition in §2), explanations, and estimations. Importantly, in Section 8, ISO articulates that the opinions generated may be expressed in a variety of ways, including as qualitative opinions or categorical opinions. When qualitative opinions are expressed, they are required to be drawn from a predefined scale of verbal expressions.

Using ISO 21043 - Part 4 parlance, this FSSB document specifically provides guidance to address the forensic-science question of source (ISO 5.1.5) using an evaluative interpretation (ISO 5.2.3) to generate an opinion in the form of a likelihood ratio (ISO 6.2) that is expressed as a qualitative opinion (ISO 8.3) from a predefined opinion scale (See Table 1 below). This FSSB guidance document also permits a categorical opinion of exclusion. As shown in Table 1, this document does not address any of the following: investigative interpretations; the estimation of any quantitative or other qualitative properties; or taxonomies for classifying organisms, objects, or patterns.

Table 1: Scope of FSSB Guidance Document for Opinion-Scale Standards for Comparative Examinations			
from ISO 21043 Forensic sciences Part 4: Interpretation			
Types of Questions	**Types of Interpretation**	**Types of Opinions**	**Expression of the Opinion**
A1. Classification, including identification: What is this item?	B1. Investigative Interpretation - No Propositions	C1. Likelihood Ratios (LR), relative probabilities of the observations given the propositions	D1. Quantitative expression of opinion
A2. Quantification: How much is there of an item or how many are there of an item?	B2. Evaluative Interpretation - At least 2 competing propositions are used	C2. Categorical Opinions on the propositions	D2. Qualitative expression of opinion *(shall be drawn from predefined opinion scale)*
A3. Question of Source: Who or what is the source of an item, or do two or more items have the same source?		C3. Explanations generated for the observations	D3. Categorical Opinions
A4. Reconstruction: What happened, where did it happen, when did it happen and who did what?		C4. Estimations	

Green: This document's guidance applies to the required combination of question, interpretation, opinion, and expression pathway (A3+B2+C1+D2).

Yellow: A limited permission is given for categorical opinions of exclusion only.

Note: The guidance document applies only to opinion scale standards being developed for comparative examinations to help answer the question of source using an evaluative interpretation that generates an opinion in the form of a likelihood ratio that is expressed qualitatively.

2. Terminology

In this document:

2.1. "**Observations**" refer to the result of analysis of items or of the scene.[1] They are the input to interpretation.

Note: Depending on the context, "evidence" refers either to the items or to the observations from the examination of the items.

2.2. "**Interpretation**" is the part of the examination that uses professional judgment, logic, expertise, and relevant data and information and if applicable, statistical models to infer the meaning of observations so as to provide opinions with respect to questions asked.[2]

2.3. **"Opinion"** is the examiner's judgment based on the interpretations of observations.[3]

2.4. "**Categorical opinion**" is an opinion that, in the examiner's judgment, one proposition or explanation is true, and it would be impossible to obtain the observations if any other alternative propositions or explanations considered were true. Example: In the examiner's judgment, the shoe did not make the questioned shoemark (i.e., exclusion).[4]

2.5. "**Probability**" is an assessment of the chance that an event will occur (or has occurred) or that a proposition (sometimes referred to as a hypothesis) is true. Probabilities can be quantitative (conventionally on a scale ranging from 0 to 1) or expressed in words to denote an interval containing a numerical probability, although the examiner may not define either the probability or the interval explicitly.

2.6. "**Prior probability**" is an assessment of the probability of an event or hypothesis reached on the basis of information known or assumed to be true before considering new evidence.

2.7. "**Posterior probability**" is the prior probability as adjusted in light of new evidence. If the prior probability of some hypothesis H_1 (such as "the reference and questioned items originated from the same source") is denoted $P(H_1)$, then the posterior probability can be denoted $P(H_1|E)$, where the vertical bar stands for "given" or "conditional on." Posterior probabilities or ratios of posterior probabilities, called posterior odds, speak to the belief in the truth of the hypotheses by their assessor.

[1] ISO 21043-1:2025 Forensic sciences — Part 1: Vocabulary

[2] Ibid

[3] Ibid

[4] Ibid

2.8. "**Strength-of-evidence**" or "**support**" is the extent to which evidence is probative of one hypothesis (H_1) as opposed to an alternative hypothesis (H_2), where the hypotheses are mutually exclusive (only one can be true).[5]

Note: "Strength-of-evidence" is a function of a "likelihood ratio." A likelihood ratio equals $P(E|H_1)$—the probability of the evidence given H_1—divided by $P(E|H_2)$—the probability of the evidence given H_2.

2.9. "**Verbal scale**" consists of words rather than (or in addition to) numbers that characterize either the assessed strength of the evidence or the probable truth of the hypotheses. Although both scale types are possible, this document only applies to strength-of-evidence scales and categorical opinions of exclusion since it is generally agreed that truth-of-hypotheses scales for inclusions are outside the purview of forensic scientists or examiners.

2.10. "**Shall**" indicates a requirement, "**should**" indicates a recommendation, "**may**" indicates a permission, and "**can**" indicates a possibility or a capability.

3. Required Sections for Evaluative Opinion-Scale Standards

Strength-of-evidence opinion-scale standards shall address the following matters:

1. Scope.
2. Normative references (including any examination-method standards that reference the opinion-scale standard).
3. Terminology.
4. The nature of the scale, including:
   a. The meaning of each level of verbal expression in the scale.
   b. Guidance on how to communicate scale information to factfinders, such as judges or jurors, since $P(E|H_1)$ can be easily misinterpreted as $P(H_1|E)$, and a likelihood ratio can be misconstrued as posterior odds. This guidance will make it more likely that factfinders interpret the findings as intended.
   c. Limitations or cautions in presenting the opinion on the selected scale (see § 9).
5. Examples of using the scale in reporting.

4. Number of Levels in Verbal Opinion Scales

1. No minimum or maximum number of levels of verbal expression is required.
2. Subcommittees shall fully describe the levels of verbal expression and their use, and consider what information supports the ability of examiners to use the number of

[5] Although, for any pair of mutually exclusive hypotheses, the likelihood ratio indicates the extent to which the evidence supports one hypothesis relative to the alternative hypothesis, the two hypotheses might omit important possibilities. If so, it is then necessary to consider likelihood ratios (or constituent likelihood values) for multiple pairs of hypotheses. See, e.g., Hicks et al. (2021); compared to Lund and Iyer (2017), and Lund and Iyer (2026).

levels of verbal expression. This information includes the basis for differentiation (resolution) between the levels of verbal expression.

5. Separation of Opinion Scales and Categorical Source Opinions

1. In general, scales for source opinions shall be for expressions of the strength-of-evidence (support). In evaluative reporting, statements about the extent to which the examiner believes source propositions to be true or false are beyond the scope of the examiner's responsibility.
2. The only time the examiner may communicate directly about the truth of a source proposition is when the examiner excludes a proposed source because the examination of the features indicates that it is impossible for the proposed source to possess or to have produced the observed features. Standards shall not use either categorical or probabilistic expressions for the truth of single-source attributions.
3. If a standard allows for both the categorical opinion of exclusion and for statements of support, it shall make it clear that the categorical opinion is not part of the scale for expressing support. If they are both depicted in a single diagram, they shall not be contiguous.

6. Performance Data

As noted in section 4, opinion scale standards are required to define the levels of verbal expression that comprise an opinion scale, which includes presenting general descriptions for the meaning of each level. To further enhance usability, objective criteria for distinguishing between the levels of verbal expression on an opinion scale shall ultimately be demonstrated through performance data (e.g., interlaboratory testing evaluating the scale's application). Performance data may be provided either in the opinion scale standard or in interpretation, method, or practice standards that discuss the interpretation of observations for specific analyses or comparisons using the scale. If performance data exist, they shall be referenced. If no such data exist when the scale is first developed, their absence shall be noted in the scale standard. Thus, subcommittees may draft a scale standard to be used in conjunction with specific examination-method standards or other documents to develop performance data.

7. Discipline-specific Language and Structure of Scale Standards

1. No one set of labels for the verbal expression of levels of support for the hypotheses in scale standards is required for all scales. Subcommittees may adopt appropriate descriptive words for strength-of-evidence or support, such as “extremely strong,” “strong,” “moderately strong,” or “weak.” Examples of strength-of-evidence and truth-of-hypothesis scales are included in the Appendix to help drafters understand the difference between these types of scales.
2. Strength-of-evidence scale standards may have a different number of levels of verbal expression above and below the category for equal support.
3. If suitability is addressed in a scale standard, the standard shall differentiate between unsuitable for analysis or interpretation (in the quantity or quality of the items) and

inconclusive or equal levels of support. Suitability is evaluated before the opinion scale is applied.

8. Presentation of the Entire Opinion Scale

Standards dealing with reporting, interpretation, and expression of opinions shall require the case report to contain a copy of the entire verbal scale. This is consistent with ISO 21043 Forensic sciences - Part 5 Reporting § 5.5.1.2, which requires reports containing opinions from an opinion scale to include a reproduction of the entire scale or a reference in the report.

9. Limitations on the Attributes and Use of Scale Standards

A section describing the limitations of the scale shall be included in the standard. Limitations should include prohibitions on the use of particular terminology that increases the potential for misinterpretation or overstatement of the findings. Examples include:

1. An examiner shall not assert that a "source identification" or a "source exclusion" is based on the "uniqueness" of a type of evidence.
2. An examiner shall not use the terms "individualize" or "individualization" for a same-source opinion.
3. An examiner shall not assert that two items originated from the same source just because their properties or features are very similar (and any differences are explicable); nor shall an examiner assert that items match "to the exclusion of all other sources."
4. An examiner shall not assert that examinations are infallible or have a zero-error rate in a forensic discipline.
5. An examiner shall not cite the number of examinations conducted in the forensic discipline during their career as a measure of the accuracy of the opinion provided. An examiner may cite the number of examinations they have conducted during their career to establish, defend, or describe their qualifications or experience.
6. An examiner shall not assert that two samples originated from the same source with absolute or 100% certainty, with a "reasonable degree of scientific certainty," "reasonable scientific certainty," or with a discipline-specific degree of reasonable certainty.

Appendix: Examples of Scales.

This appendix contains examples of (1) strength-of-evidence scales and (2) truth-of-hypotheses scales. The examples are chosen because they use one or the other mode when articulating opinions exclusively. Although they are or have been used in forensic-science practice or proficiency testing, they are included here solely to display the difference between the two modes of articulating opinions, and not to endorse any particular terms, level of verbal expression, or definitions.

1. Strength-of-Evidence Opinion Scales

Example 1

The following strength-of-evidence opinion scale does not require an examiner to assign a number to their likelihood ratio, it describes observations that give rise to the ratio, which is presented in terms of support for the hypothesis. It is adapted from *ISO 21043 Forensic Sciences - Part 4: Interpretation* - Annex B, Table B.2.

Verbal expression	***Description of observations***
Strongest support for same source	A lot of similarity and no unexplainable difference can be observed in the comparison. The examiner would not expect to make the observations in an item from a different source.
Strong support for same source	Similarity and no unexplainable difference can be observed in the comparison. The examiner has a much higher expectation of making the observations in an item from the same source than in an item from a different source.
Limited support for same source	A little similarity and no unexplainable difference can be observed in the comparison. The examiner has a somewhat higher expectation of making the observations in an item from the same source than in an item from a different source.
No support (for either proposition)	The observations do not provide support for one proposition over the other. The examiner's expectation for the observations of the comparison is similar for an item from the same or from a different source. This expression is not meant to be used when no comparison could be carried out.
Limited support for different source	A little difference and no specific similarity can be observed in the comparison. The examiner has a somewhat higher expectation of making the observations in an item from a different source than in an item from the same source.
Strong support for different source	Difference and no specific similarity can be observed in the comparison. The examiner has a much higher expectation of making the observations in an item from a different source than in an item from the same source.
Strongest support for different source	A lot of difference (or any unexplainable difference) and no specific similarity can be observed in the comparison. The examiner would not expect to make the observations in an item from the same source.

Example 2

The following scale applies levels of verbal expression to numerical likelihood ratios (LRs). It is adapted from *ISO 21043 Forensic Sciences - Part 4: Interpretation* - Annex B, Table B.1. The LR expresses the examiner's assessment of the relative probabilities of the observations if proposition A is true versus proposition B is true. In the scale below, the observations were more probable if proposition A were true than if proposition B were true. If the opposite were true, the order of proposition A and proposition B in the verbal expression would be reversed.

While this guidance document is focused on opinion scales where the strength of the opinion is expressed qualitatively, it is important to note that in examinations where a numerical likelihood ratio is calculated using quantitative observations, quantitative data, and a statistical model, the opinion should be expressed using that numerical value.[6]

It is important to note, if an interpretation method does not produce a numerical value (which again is the primary focus of this document), a verbal expression of the qualitative opinion shall be recorded. As a matter of practice, opinion scales shall not be used to convert a qualitative verbal expression of an opinion into a numerical value.[7]
Note: The # represents discipline-specific numerical values.

Range of the likelihood ratios (LR)	***Verbal expression of strength of opinion***
# < LR < #	The observations are approximately equally probable irrespective of whether proposition A were true or proposition B were true. The observations do not support one proposition over the other.
# ≤ LR < #	The observations are slightly more probable if proposition A were true than if proposition B were true. The observations provide weak support for proposition A relative to proposition B.
# ≤ LR < #	The observations are more probable if proposition A were true than if proposition B were true. The observations provide moderate support for proposition A relative to proposition B.
# ≤ LR < #	The observations are appreciably more probable if proposition A were true than if proposition B were true. The observations provide moderately strong support for proposition A relative to proposition B.
# ≤ LR < #	The observations are much more probable if proposition A were true than if proposition B were true. The observations provide strong support for proposition A relative to proposition B.
# ≤ LR < #	The observations are far more probable if proposition A were true than if proposition B were true. The observations provide very strong support for proposition A relative to proposition B.
LR ≥ #	The observations are exceedingly more probable if proposition A were true than if proposition B were true. The observations provide extremely strong support for proposition A relative to proposition B.

[6] ISO 21043-4:2025 Forensic Sciences - Part 4: Interpretation
[7] Ibid

2. Truth-of-Hypothesis Scales

Opinions expressed on truth-of-hypothesis scales present degrees of belief in source hypotheses. This guidance document covers strength-of-evidence rather than truth-of-hypothesis scales for evaluative attribution opinions. Solely for the purpose of clarifying the difference between the types of scales, this section gives examples of truth-of-hypothesis ales. It does not endorse the use of these truth-of-hypothesis scales.

Example 1

The following "five-level conclusion scale" for handwriting comparisons was studied in Hicklin et al. (2022). It is a verbal scale for posterior-probability judgments.

- The questioned sample was written by the known writer
- The questioned sample was probably written by the known writer
- No conclusion
- The questioned sample was probably not written by the known writer
- The questioned sample was not written by the known writer

Example 2

The following scale uses verbal tags for defined ranges of a "probability of paternity." It is adapted from Table 8 of Bryant (1988). The "probability of paternity" is a posterior probability that presumes that the probability of paternity prior to considering the genetic data is one-half.

Probability	*Verbal Predicate*
99.8% to 99.9%	Practically proven
99.0% to 99.7%	Extremely likely
95.0% to 98.9%	Very likely
90.0% to 94.9%	Likely
80.0% to 89.9%	Hint
Less than 80%	Not useful

Another set of "verbal predicates" associated with reported paternity probabilities comes from Charles Brenner, *Hummel's Verbal Predicates*, https://dna-view.com/hummel.htm (citing Part 1 of the same source). Although extensively criticized in the legal and statistical literature, these probabilities of paternity continue to be reported and admitted into evidence in civil and criminal cases.

Probability	*Verbal Predicate*	*Likelihood Ratio*
(99.9%)	practically proven	
99.8%		399
	highly likely	
99%		99
	very likely	
95%		19
	likely	
90%		9
(50%)	(no verbal predicate)	
10%		1/9
	unlikely	
5%		1/19
	very unlikely	
1%		1/99
	highly unlikely	
0.2%		1/399
(0.1%)	practically excluded	

Bibliography

1) The writings listed below contain explanations or recommendations for presenting interpretations of data acquired in forensic examinations. They are a small sample of books, articles, reports, position statements, and recommendations on the subject.
2) Aitken, Colin G G, F Taroni & S Bozza (2024). The Role of the Bayes Factor in the Evaluation of Evidence. *Annual Review of Statistics and its Application* 11: 203–226. https://doi.org/10.1146/annurev-statistics-040522-101020
3) Aitken, Colin G G, F Taroni & S Bozza (2021). *Statistics and the Evaluation of Evidence for Forensic Scientists*. Chichester, UK: John Wiley & Sons. 3rd ed.
4) Aitken, Colin G G & 30 coauthors (2011). Expressing Evaluative Opinions: A Position Statement, *Science & Justice* 51: 1, http://dx.doi.org/10.1016/j.scijus.2011.01.002
5) Aitken, Colin G G, et al. (2010). *Fundamentals of Probability and Statistical Evidence in Criminal Proceedings: Guidance for Judges, Lawyers, Forensic Scientists and Expert Witnesses*, http://www.rss.org.uk/Images/PDF/influencing-change/rss-fundamentals-probability-statistical-evidence.pdf [https://perma.cc/NV7K-VJ9C] (committee of the Royal Statistical Society)
6) *American Statistical Association Position on Statistical Statements for Forensic Evidence* (2019). https://www.amstat.org/asa/files/pdfs/POL-ForensicScience.pdf [https://perma.cc/X4AM-AVBU]
7) Banks, David, K Kafadar, D H Kaye & M Tackett (eds. 2021). *Handbook of Forensic Statistics*. Boca Raton, FL: Chapman & Hall/CRC Press (chs. 1, 4 and 10)
8) Berger, Charles E.H. (2025). Finally a Really Forensic Worldwide Standard: ISO 21043 Forensic Sciences, Part 4, Interpretation, *Forensic Science International: Synergy*, 10: 100589, https://doi.org/10.1016/j.fsisyn.2025.100589
9) Biedermann, Alex, W C Thompson, & J Vuille (2018). After Uniqueness: The Evolution of Forensic-Science Opinions, *Judicature* 102(1): 18–27
10) European Network of Forensic Science Institutes (2015). *ENFSI Guideline for Evaluative Reporting in Forensic Science*, http://enfsi.eu/wp-content/uploads/2016/09/m1_guideline.pdf [https://perma.cc/H296-YKML] (Sheila M. Willis, Louise McKenna, L., Sean McDermott, Geraldine O'Donell, Aurélie Barrett, Birgitta Rasmusson, Anders Nordgaard, Charles E.H. Berger, Marjan J. Sjerps, José Juan Lucena-Molina, Grzegorz Zadora, Colin G.G. Aitken, Luan Lunt, Christophe Champod, Alex Biedermann, Tasha N. Hicks, Franco Taroni)
11) Expert Working Group (2024). *Forensic DNA Interpretation and Human Factors: Improving Practice Through a Systems Approach*. Gaithersburg, MD: National Institute of Standards and Technology
12) Expert Working Group (2012). *Latent Print Examination and Human Factors: Improving the Practice Through a Systems Approach*. Gaithersburg, MD: National Institute of Standards and Technology
13) Forensic Science Regulator, FSR-C-118 *Development of Evaluative Opinions* (2021). https://www.gov.uk/government/publications/development-of-evaluative-opinions
14) Good, I J (1991). Weight of Evidence and the Bayesian Likelihood Ratio. In *The Use of Statistics in Forensic Science*, Aitken, Colin G G & D A Stoney (eds.), pp. 85–106
15) Hicklin, R. Austin, Linda Eisenhart, Nicole Richetelli, Meredith D. Miller, Peter Belcastro, Ted M. Burkes, Connie L. Parks, Michael A. Smith, JoAnn Buscaglia, Eugene M. Peters, Rebecca Schwartz Perlman, Jocelyn V. Abonamah, & Brian A. Eckenrode (2022), Accuracy

and Reliability of Forensic Handwriting Comparisons, Proceedings of the National Academy of Science, 119(32): e2119944119 (2022), https://doi.org/10.1073/pnas.2119944119

16) Hicks, Tacha et al., Comparing multiple POI to DNA mixtures, Forensic Science International: Genetics, Volume 52, 2021, 102481, https://doi.org/10.1016/j.fsigen.2021.102481

17) Kaye, David H (2012). Likelihoodism, Bayesianism, and a Pair of Shoes, *Jurimetrics: The Journal of Law, Science, and Technology*, 53(1): 1–9, http://ssrn.com/abstract=2269088

18) Kaye, David H, et al. (2022). Toolmark-Comparison Testimony: A Report to the Texas Forensic Science Commission, available at http://ssrn.com/abstract=4108012 and in the Commission's *Final Report on Complaint No. 21.27*, Apr. 26, 2024, at https://www.txcourts.gov/media/1459122/complaint-2127-final-report-09042024.pdf

19) Kaye, David H, D Bernstein & J Mnookin (2011). *The New Wigmore, A Treatise on Evidence: Expert Evidence* (2d ed.) New York, NY: Aspen Pub. (chs. 13–15)

20) Lund, Steven P. & Iyer, Hari (2017). Likelihood Ratio as Weight of Forensic Evidence: A Closer Look, *Journal of Research of National Institute of Standards and Technology,* 122:27. https://doi.org/10.6028/jres.122.027

21) Lund and Iyer, "The Influence of Validation Data on Logical and Scientific Interpretations of Forensic Expert Opinions", (in press) *Law, Probability, and Risk*, Volume 25, Issue 1, January 2026, https://doi.org/10.1093/lpr/mgaf013

22) National Institute of Forensic Sciences of Australia and New Zealand (2017). *An Introductory Guide to Evaluative Reporting*, https://www.anzpaa.org.au/forensic-science/our-work/products/publications

23) Neumann, Cedric, Kaye, D, Jackson, G, Reyna V, & Ranadive, A (2016). Presenting Quantitative and Qualitative Information on Forensic Science Evidence in the Courtroom, *Chance* 29(1): 37-43

24) Nordgaard, Anders & B Rasmusson (2012). The Likelihood Ratio as Value of Evidence—More than a Question of Numbers, *Law, Probability & Risk* 11(4): 303-315

25) Nordgaard, Anders, R Ansell, D Weine, J & Jaeger (2012). Scale of Conclusions for the Value of Evidence, *Law, Probability & Risk* (2012). 11(1): 1–24

26) Samie, Lydie et al. (2025). Propositions Used to Assess the Value of Forensic DNA Mixtures in an FSWG-ISFG Interlaboratory Comparison, *Forensic Science International: Genetics*, 78: 103281, https://www.fsigenetics.com/article/S1872-4973(25)00061-4/fulltext

27) Texas Forensic Science Commission (2024). *Final Report on Complaint No. 21.27*, pp. 46–51, https://www.txcourts.gov/media/1459122/complaint-2127-final-report-09042024.pdf

28) Wivell, Richard et al. An Investigation into Compound Likelihood Ratios for Forensic DNA Mixtures, *Genes* (2023).14(3):714, doi:10.3390/genes14030714