



# NATIONAL COMMISSION ON FORENSIC SCIENCE

**NIST**  
National Institute of  
Standards and Technology  
U.S. Department of Commerce

## Views of the Commission Statistical Statements in Forensic Testimony

---

<b>Subcommittee</b>
Reporting and Testimony
<b>Status</b>
Initial Draft

<b>Date of Current Version</b>	13/11/16
<b>Approved by Subcommittee</b>	16/11/16
<b>Approved by Commission</b>	[dd/mm/yy]

*Note: This document reflects the views of the National Commission on Forensic Science and does not necessarily represent the views of the Department of Justice or the National Institute of Standards and Technology. The portion of the document directly labeled “Views of The Commission” represents the formal Views of the Commission. Information beyond that section is provided for context. Views documents do not request specific action by the Attorney General, and thus do not require further action by the Department of Justice upon their approval by the Commission. The National Commission on Forensic Science is a Federal Advisory Committee established by the Department of Justice. For more information, please visit: <https://www.justice.gov/ncfs>.*

**COMMENT:**

by

*Geoffrey Stewart Morrison*

Independent Forensic Consultant

Adjunct Professor, University of Alberta

<http://geoff-morrison.net/>

<http://forensic-evaluation.net/>

Version 2017-01-04a

I am grateful to the subcommittee and the working group for drafting the document on statistical statements in forensic testimony. I am very supportive of what I understand to be its goals.

My understanding of the 2016-11-16 draft is that the subcommittee recommends that:

- Strength of evidence should be evaluated and expressed as a likelihood ratio.
- Likelihood ratios should be calculated on the basis of relevant data and statistical models.
- The accuracy and precision of the procedures and systems used to calculate likelihood ratios should be empirically assessed.

My aim in making these comments and recommended revisions is to assist the subcommittee to express these ideas more clearly. I hope they will be recognized as friendly suggestions. My intention was to be faithful to what I understood to be the core ideas that the committee wanted to express, but in some parts of the document my thoughts on these matters may differ from those of the subcommittee and my suggested revisions may have ended up too different from what the subcommittee wanted to say.

Below, between paragraphs from the original document I have inserted comments and recommended revisions. A complete redraft of the whole document including all the recommended revisions is also presented at the end of the present document.

I would be happy to provide additional advice. If there is anything in my comments which is unclear, perhaps a teleconference would be the most efficient way to clarify. A recent paper that might help in understanding my thoughts on this topic is Morrison & Thompson (2017).

## Overview

This Views document presents background information and views on the following question: When experts present the results of forensic science and medicine examinations, tests, or measurements in reports or testimony, what types of quantitative or qualitative statements should they provide to indicate the accuracy of measurements or observations and the significance of these findings? This document refers to such statements as “statistical statements.” These statistical statements may describe measurement accuracy (or conversely, measurement uncertainty), weight of evidence (the extent to which measurements or observations support particular conclusions), or the probability or certainty of the conclusions themselves. Such statements occur with many types of forensic evidence. Five examples follow.

### COMMENT:

The opening paragraph mixes evaluation of strength of evidence and validation of performance. These are both important, but separate steps. Confusion between the two is apparent in the PCAST report and may have arisen from thinking in a “match” / “non-match” framework. Relevant data are necessary to train statistical models in order to calculate likelihood ratios, but a second set of test data is then necessary to test the performance of the system trained on the first set of data. It is well established that one should train and test on separate data since training and testing on the same data will lead to results which are overly optimistic with respect to how the system will perform when it is actually applied to make inferences about new data. Either completely separate sets of data or cross validation should be used to avoid training and testing on the same data. The PCAST report discussed *sensitivity* and *false-alarm rates* when a “match” is declared. *Sensitivity* and *false-alarm rates* look like measures of performance, but in fact they were being used to quantify the strength of the evidence: *sensitivity* divided by *false-alarm rate* gives a likelihood ratio. The values for *sensitivity* and *false-alarm rate*, and hence the value for the likelihood ratio, are calculated using a set of training data, a separate set of test data should subsequently be used to test the performance of the system. The confusion between evaluation of strength of evidence and validation of performance would be less likely to occur if the two tasks were addressed separately, and the task of calculating the strength of evidence were not conceptualized in terms of “match” / “non-match” and *sensitivity* and *false-alarm rate* / *correct-rejection* and *miss rate*. For data that are legitimately discrete, the appropriate conceptualization would be: the probability of the evidence if the same-origin hypothesis were true divided by the probability of the evidence if the different-origin hypothesis were true. For data that are continuous, the appropriate conceptualization would be: the likelihood of the evidence if the same-origin hypothesis were true divided by the likelihood of the evidence if the different-origin hypothesis were true. Continuous data should not be dichotomized, see Morrison, Kaye, et al (2016). I therefore recommend that the document be divided into two parts, one addressing evaluation of strength of evidence, and the other addressing validation of systems which

evaluate strength of evidence.

I have already above simplified the discussion to consideration of same-origin versus different-origin hypotheses. This appears to be the focus that the 2016-11-16 draft took when it stated later that “the discussion that follows is of special relevance to pattern, impression, and trace evidence.” I believe that the final document will be much more effective if it is more focused and clearly addressing these hypotheses from the start. I think that it is helpful to include extensions to examples such as above versus below a legally prescribed threshold, and simple activity hypotheses which are amenable to empirical investigation, but think it would be counterproductive to attempt to reach further. In particular, I think that it will be more effective and less confusing if the current document is limited to trying to improve the situation with respect to statistical evaluation in forensic science, and that forensic medicine be addressed separately. I think that there are sociocultural differences between the two fields and potentially more complex issues to be address in the latter. If progress is first made in forensic science, arguments and examples can then be extended to forensic medicine.

Some of the specific wording in the opening paragraph I find problematic:

“significance of these findings” appears to refer to what I would call “strength of evidence”, and I assume was not intended to be a reference to frequentist null-hypothesis testing. I recommend avoiding the use of the word “significance” completely, except if used with its technical statistics meaning, and since the latter is not relevant for evaluation of strength of evidence, that it not be used at all in this document.

“accuracy of measurements” If the focus of the document is evaluation of strength of evidence, then it is the accuracy of the system that evaluates strength of evidence that is most important to asses, rather than accuracy of the measurements which are a component of the system. Measurements could be extremely accurate, but if the measurements contain no useful information for addressing the hypotheses to be evaluated, then that accuracy will be irrelevant.

“measurement accuracy (or conversely, measurement uncertainty)” This appears to be confusing accuracy and precision.

“the probability or certainty of the conclusions themselves” It is unclear to me what is meant here. “probability” and “certainty” are not, for me, synonyms.

“expert” I believe that the preferred term here would be “practitioner” since the latter is more broadly applicable whereas the former is someone who has been qualified by the court to give expert evidence. At the time the practitioner is performing the analysis they will not yet have been qualified as an expert witness in the court case. The case may be resolved pre-trial or for other reasons they may not be called to give evidence and hence the practitioner may never be qualified as an expert in the case.

## RECOMMENDED REVISION:

### Overview

This Views document presents background information and views on the following question: How should forensic practitioners evaluate and report *strength of evidence*? [FN: “strength of evidence” has also been referred to as “weight of evidence” and as “probative value”] The statements of strength of evidence that forensic practitioners make in reports and testimony are referred to in this document as *statistical statements*. Such statements are made in relation to many types of forensic evidence. Five examples follow.

1. *Pattern and impression evidence*. A shoe print found in the dirt next to the deceased is compared with a print made from the suspect’s sneaker. What degree of similarity exists between the two impressions, and how strongly does it support a claim that the suspect’s sneaker is, or is not, the source of the print in the dirt?

#### COMMENT:

Since both similarity (in numerator of likelihood ratio) and typicality with respect to the relevant population (in denominator of likelihood ratio) need to be considered, I recommend against referring only to similarity.

In the recommended revision, I make more explicit that fact that two hypotheses have to be considered. I also change “suspect” to “defendant” since I explicitly refer to “defense”.

I substituted “at a crime scene” for “in the dirt next to the deceased”. I think it best to avoid unnecessary elaboration. Also, to reduce the potential for cognitive bias, the forensic practitioner should not be exposed to task irrelevant information – “dirt” would be relevant, but “deceased” not.

#### RECOMMENDED REVISION:

1. *Pattern and impression evidence.* A shoe print found at a crime scene is compared with the sole of the defendant’s shoe. How strongly do the results of the examination support the prosecution’s contention that the defendant’s shoe is the source of the crime-scene print versus the defense’s contention that it is not?

2. *Trace evidence.* A burglar smashed a pane of window glass to open a window. Various physical and chemical properties of a glass fragment collected from the suspect’s clothing and a fragment known to be from the pane are measured. How accurate are the measurements, and how strongly does the similarity between the two sets of measurements support or refute the claim that the broken pane is the source of the fragment on the suspect as opposed to a claim that the fragments are from a different source?

#### COMMENTS:

There would normally be multiple fragments available for analysis, and multiple fragments are useful for estimating within-source variability. In the recommended revision I have therefore made it fragments of glass, plural.

For the reasons explained above, I removed reference to accuracy of measurements.

Again I removed reference to similarity alone. I added “and compared”.

I removed “or refute” as confusing. Support one hypothesis versus support another hypothesis was already the structure. “Support or refute” may suggest only one hypothesis is going to be considered and the evidence may point in favor or against that one hypothesis.

To avoid unnecessary elaboration, I simplify “smashed a pane of window glass to open a window” to “smashed a window”.

#### RECOMMENDED REVISION:

2. *Trace evidence.* A burglar smashed a window. Various physical and chemical properties of fragments of glass collected from the suspect’s clothing and fragments from the broken window at the crime scene are measured and compared. How strongly do the results of the analysis support the prosecution’s contention that the broken window at the crime scene is the source of the fragments found on the suspect’s clothing as opposed to the suspect’s claim that the fragments are from a different source?

#### COMMENTS:

Given that a glass evidence example has just been presented, this is an ideal opportunity to introduce an activity-level example.

#### RECOMMENDED REVISION:

3. *Activity level.* A burglar smashes a window. Four hours later a suspect is arrested. Two fragments of glass are found in the suspect's jacket pocket. The suspect denies involvement in the burglary, but says he did accidentally break a window in the same building two weeks earlier (he works as a window cleaner). How probable is it that the suspect would have two fragments of glass in his pocket if he had broken a window four hours earlier and two weeks earlier versus if he had only broken a window two weeks earlier?

3. *Qualitative analysis.* An oily substance is analyzed with thin layer chromatography to ascertain whether it contains tetrahydrocannabinol (THC) and hence is liquid cannabis. How sensitive and specific is the procedure for detecting THC?

#### COMMENT:

There may be circumstances in which this example is relevant (e.g., liquid cannabis could be used to poison someone), but given that cannabis in general is legal in an increasing number of jurisdictions, this example may raise distracting questions. In the suggested revision, I substitute an example involving ignitable liquid residue.

The title "qualitative analysis" appears to be inappropriate. Was "quantitative" intended?

As explained above, I think it best to avoid discussion of strength of evidence in terms such as *sensitivity* and *specificity*.

#### RECOMMENDED REVISION:

4. *Quantitative analysis.* A chemical analysis of fire debris is conducted in an arson investigation. Results are found that are thought to be positive indicators for the presence of ignitable liquid residue. What is the probability of getting these results if ignitable liquid really had been present before the fire started versus if ignitable liquid had not been present (positive results have been known to occur for other reasons).

4. *Quantitative analysis with extrapolation.* A blood sample is collected from a driver 2 hours after an accident. A chemical test indicates a blood alcohol concentration (BAC) of 0.04%, which is below the 0.08% legal limit. What is the standard error of measurement? How likely is it that the true BAC at the time of the accident was above 0.08%, given that the measurement was 0.04% 2 hours later? What is the probability of observing 0.04% at the time of the accident if the true BAC at that time was below the legal limit?

#### COMMENT:

For the reasons explained above, I removed reference to standard error of measurements.

The last two sentences were worded inconsistently. In the recommended revision I use parallel wording to emphasize the parallel calculations needed.

RECOMMENDED REVISION:

5. *Quantitative analysis with extrapolation.* A blood sample is collected from a driver 2 hours after an accident. A chemical test indicates a blood alcohol concentration (BAC) of 0.04%, which is below the 0.08% legal limit. How probable is it that the BAC at the time of the accident was above the legal limit? How probable is it that the BAC at the time of the accident was below the legal limit?

5. *Cause, manner, and time of death.* An autopsy reveals injuries believed to be indicators of child abuse. Can we infer that child abuse had occurred and caused the injuries and death?<sup>1</sup>

---

<sup>1</sup> Inferring causation is usually a statistical activity. Scientists usually infer from a treatment or cause to the outcome or effect, and one of the strongest forms of evidence to support such inference comes from randomized experiment. In the present context, we are in effect reversing that process and attempting to infer the cause from the effect, as more often is the case in the context of the law.

COMMENT:

For reasons explained above, I recommend deleting this example. The example itself raises all sorts of questions which would take a long time to address.

These statistical statements and those that appear in connection with many other forms of evidence should be based on: (1) the existence of a relevant database describing characteristics, images, observed data, or experimental results; (2) a statistical model that accurately assesses the strength of the inference in question or describes the process that gives rise to the data linked to the question at hand; (3) information on variability and errors in measurements or in statistics or inferences derived from measurements; and (4) a statistical statement regarding the probative value of any comparisons done or calculations performed (e.g., how rare is an observed positive association when two items arise from the same source and when they arise from different sources?).

COMMENT:

In the recommended revision I first simplify, and later fill in the details specific to a same-source / different-source scenario. The original gradually transitioned into the latter scenario. I make it more explicit.

I think (1) is unnecessarily complicated, “relevant data” should encompass what needs to be said here. “database” is more specific than “data” and relevant data may or may not come from a database.

In (2), whether a statistical model produced accurate results is a question for validation. I have inserted “appropriate”.

Point (3) relates to validation, and I have dropped it here.

I expand and make point (4) more explicit.

I move another paragraph from later in the document, modify it, and insert it here. The original read:

The National Research Council Committee on Identifying the Needs of the Forensic Sciences Community emphasized the importance of describing uncertainties in measurements and inferences.<sup>2</sup> Statistics is concerned with the study of variability, uncertainty, and decision-making in the face of uncertainty. It supplies a set of principles, based on probability, for drawing conclusions from data and for expressing the risks of certain types of errors in measurements and conclusions. This framework applies throughout forensic science and medicine, but the discussion that follows is of special relevance to pattern, impression, and trace evidence.

---

<sup>2</sup>Comm. on Identifying the Needs of the Forensic Sci. Cmty., Nat'l Research Council, Strengthening Forensic Science in the United States: A Path Forward 184 (2009).

This paragraph mixes evaluation of evidence and validation of performance. I have made it exclusively about the former by using only modified versions of the second and third sentence. Most of the revised version of this paragraph appears as the second paragraph below, the paragraph beginning “Statistics is concerned with”. A modified version of the last sentence of the original paragraph becomes the first sentence of the fourth paragraph below, beginning “The framework we describe here”.

I also move another paragraph from later in the document and insert it here. That paragraph begins “For many types of evidence,” and appears as the third paragraph below. I have not changed its content (apart from changing “forensic science experts” to “forensic practitioners” and “mathematical” to “statistical”, the latter for consistency). I think this paragraph is really a preliminary comment, and that it is better to get it out of the way before going into details about statistical statements.

I also move forward and modify half a later paragraph. I center the revised version on a shoe print example which I think is intuitive.

But a “positive association” is not probative unless it is more probable when the items have a common source than when they originate from different sources. Indicating the statistical weight of the positive association therefore requires a statement of how common or rare the association is, based on a database linked to the case at hand. For example, a positive association for the presence or absence of pigment in a hair cuticle is some evidence that the hairs have a common origin, but the significance of this association is unknown without data from relevant populations.

I explicitly introduce the term “likelihood ratio”. For some reason, the term was absent from the 2016-11-16 draft although it was obvious that this is what was being discussed. Introducing the term provides previously uninformed readers with a search term that will lead them more quickly to a large body of literature on the topic. If “likelihood ratio” was deliberately not used because of fear of existing prejudice against the term, I suggest that this is an ill-advised strategy since it may backfire if readers feel that there was a deliberate attempt to hide relevant information – this appears to have been one of the problems in the England & Wales Court of Appeal ruling in the case of *R v T*, which has been widely commented on in both forensic science and law journals.

#### RECOMMENDED REVISION:

These statistical statements and those that appear in connection with many other forms of evidence should be based on *relevant data* and *appropriate statistical models*.

Statistics is concerned with the study of variability, probability, and decision-making in the face of uncertainty. It supplies a set of principles, based on logic, for drawing inferences from data.

For many types of evidence, forensic practitioners may not currently be making statistical assessments explicitly, but they may nevertheless be presenting their findings in a manner that connotes a statistical assessment. For example, the statement that “the latent print comes from the defendant’s thumb” or “it is unlikely that the print came from anyone else” suggests a high probability that the print came from the defendant as determined by an understanding of the frequencies of similar features in fingerprints from the same individual and in prints from different individuals. As forensic science moves forward, the Commission anticipates efforts to make the presentation of analyses more overtly statistical and quantitative, and it is toward this end that the Commission expresses its views. The Commission advances a preference for an explicit statistical foundation for statements because a statistical analysis tends to provide a ready means for assessing and expressing uncertainty.

The framework we describe here applies throughout forensic science (including all of the examples listed above and more), but for simplicity and concreteness we focus on comparison of a specimen of questioned origin and a sample of known origin, e.g., a shoe print at a crime scene compared with a defendant’s shoe, or fragments of glass recovered from a suspect’s clothing compared with fragments of glass from a broken window at a crime scene (see examples 1 and 2 above). In such scenarios the question to be answered has two parts:

1. What is the probability of the observed properties of the questioned-source specimen if it came from the known source?
2. What is the probability of the observed properties of the questioned-source specimen if it came not from the known source but from some other source in the relevant population?

Part 1 can be considered a question about how *similar* the questioned-source specimen is to the known source, and part 2 can be considered a question about how *typical* the questioned-source specimen is with respect to the relevant population. [FN: More generally, parts 1 and 2 are often referred to as the *prosecution hypothesis* and the *defense hypothesis* (or the *alternative hypothesis*) respectively. The term “defense hypothesis” does not imply that the defense must advance an alternative hypothesis, but in calculating strength of evidence the forensic practitioner must consider an alternative hypothesis that is some meaningful negation of the prosecution hypothesis. The two hypotheses must be mutually exclusive and, within reason, exhaustive.] Dividing the answer to part 1 by the answer to part 2 produces a *likelihood ratio*, which is a statistical statement of the strength of evidence. [FN: A substantial body of literature exists addressing theoretical aspects of the likelihood ratio framework and describing and evaluating statistical models for calculating likelihood ratios as quantifications of strength of evidence. A new edition of a classic 1995 book-length introduction to the likelihood ratio framework is Robertson et al. (2016).]

Note that just knowing about similarity is not particularly informative as to strength of evidence. For example, if the shoe print from the crime scene and the defendant’s shoe are both the same size, they are maximally similar on this metric. If the relevant population were shoes worn by adult males in the United States, however, the strength of evidence would be relatively weak if they were both size 10, a very typical shoe size, but relatively strong if they were both size 14, a quite atypical shoe size.

*Relevant numeric data* will consist of *quantitative measurements* made on (a) the questioned-source specimen, (b) a sample of the known source, and (c) a sample of the relevant population.

At the core of all of such statistical calculations, there must be data from a *relevant population or sampling*. Impressions of the soles of shoes gathered in Israel may not be relevant to a case involving a shoe print in Alaska. If a clear population of relevance can be defined, then one needs to consider the extent to which the actual data in the database represent the population. To be applicable to casework, empirical studies of the reliability and accuracy of examiners’ judgments must involve materials and comparisons that are representative of actual cases and rely on data

from a relevant population or sample base.

#### COMMENT:

I have moved this paragraph earlier to be able to address the question of what constitutes a relevant population immediately after it is introduced. I introduced it to make the content of the previous paragraph more explicit.

In the recommended revision, I reworded the paragraph to flow better and be more explicit in its new position. I tie the examples of relevant population more closely to the previously introduced examples.

A substantial change is the deletion of “reliability and accuracy” and “examiner’s judgments”, which will be addressed later. I keep the concept of “representative of actual cases” but with reference to training data rather than test data.

I deliberately use the less formal term “echoey”, rather than “reverberant”, on the assumption that the former will be more widely understood.

#### RECOMMENDED REVISION:

The *relevant population* is the population from which the questioned-source specimen could have come if it had not come from the known source. For example, if the questioned-source specimen is a footwear print in the snow in Alaska in winter, the relevant population would at least be restricted to footwear worn in Alaska when there is snow on the ground in winter. Footwear worn in Hawai’i at any time of year would obviously not be the relevant population, but footwear worn in Alaska in summer would also not be the relevant population. Footwear worn in the city may not be the relevant population if the crime were committed in a rural area, etc., etc. If the questioned-source specimen is fragments of glass found on the suspect’s clothing, and the suspect claims to have innocently broken another window in the same building as the crime scene, then ideally it would be glass from the particular other window that would be the relevant population. If the suspect makes no specific claims as to where the glass fragments may have come from, then the relevant population could be quite wide, including window glass in that part of the world and also automotive and container glass. What constitutes the relevant population in a given case may require substantial consideration.

Since the calculated strength of evidence will depend in part on typicality with respect to the relevant population, it is important that the forensic practitioner clearly communicate what they have specified as the relevant population. The judge at an admissibility hearing and/or the trier of fact at a trial needs to know what relevant population the forensic practitioner specified in the typicality part of the question, first so that they can decide whether the forensic practitioner is actually answering a relevant question for them, and second so that they can understand the answer that the forensic practitioner provides to that question – if one does not understand the question, one cannot understand the answer.

In addition to specifying the relevant population they intend to use, the forensic practitioner must use a *sample* which is sufficiently representative of the relevant population. A sample which is too small or which is systematically biased will lead to a poor estimate of the probability of the observed properties of the questioned-source specimen if it came not from the known source but from some other source in the relevant population. This in turn will lead to a poor estimate of the likelihood ratio as a statistical statement of the strength of evidence. It may be impossible to obtain a sample that is perfectly representative of exactly the right population. The issue is whether the sample is sufficiently representative that the results of the analysis will be reasonable answers to the question asked. For example, in the broken window example, if glass fragments from the specific window that the suspect claims to have broken two weeks earlier are not available, samples of glass from other windows in the same building might be a reasonable practical alternative that implies a different but still meaningful question. Again, it is important that the forensic practitioner clearly communicate what they did so that their decisions and actions can be reviewed by the court.

In many branches of forensic science the *conditions* of the questioned specimen and known sample are relevant. For example, in a forensic voice comparison case the questioned-speaker recording may be an

intercept of a lively mobile telephone conversation on which there is background traffic noise, and the known-speaker recording may be of subdued answers to police interview questions recorded in a small echoey room. Calculating a likelihood ratio based on data that reflect conditions substantially different from those of the questioned specimen and known sample may produce misleading results, hence the forensic practitioner will need to use data which are sufficiently reflective of the conditions of the questioned specimen and known sample in the case. Again, it is important that the forensic practitioner clearly communicate what they did so that their decisions and actions can be reviewed by the court.

#### COMMENT:

There follow a number of paragraphs which mix discussion of statistical evaluation of strength of evidence, validation of performance, and human judgment. I find it increasingly difficult to disentangle the threads on a paragraph by paragraph basis, even allowing for reordering of paragraphs. I therefore make some between-paragraph comments on particular issues, but provide recommended revisions for all these paragraphs later as a single block.

The recommended revisions are structured so that they first deal with statistical models for evaluation of forensic evidence, then discuss validation, then discuss human-judgment based approaches.

Trace, impression, or pattern evidence examiners should follow a valid and reliable process to determine whether there is a positive or negative association between the item in question (often called a “questioned” sample or specimen) and a sample whose source is known (such as a reference sample from the defendant). Reliability and external validity should be established via scientific studies that have been the subject of independent scientific scrutiny.<sup>3</sup> Only when the reliability and validity of the process have been studied quantitatively can a probabilistic or statistical model for indicating the uncertainty in measurements and inferences be credible.

---

<sup>3</sup> Nat’l Comm’n on Forensic Science, Views Document on Technical Merit Evaluation of Forensic Science Methods and Practices, June 21, 2016, <https://www.justice.gov/ncfs/file/881796/download>.

Such models are most convincing when a scientific understanding of the physical process that generates the features exists. Sufficient knowledge of the process permits a mathematical model to be developed. This approach has been successful for determining the probability that associations in particular DNA features will exist among different individuals. For other types of trace and pattern evidence, however, no widely accepted probabilistic models of the phenomena that give rise to the features are available. Consequently, most efforts to provide probabilistic statements about features and their degree of association often rest on the personal impressions of examiners, supported by their subjective judgment developed through individual training and experience, or by reference to empirical studies of the reliability of the judgments of examiners. Training and experience are important in applying valid techniques, but they are not a sufficient basis for establishing the uncertainty in measurements or inferences.

#### COMMENT:

“positive or negative association” are not defined statistical terms. They may be alternative for “match / non-match” (this appears to be confirmed two paragraphs below). This approach has been critiqued elsewhere.

“reference sample” is an ambiguous term. I avoid using it.

I may disagree with some of the statements above (end of first paragraph and beginning of second paragraph). Theoretical knowledge of underlying processes may be helpful, and in some cases such as DNA modeling may even be essential, but they are not always essential. In a field such as automatic speaker recognition, which has been exploited in forensic voice comparison, much progress has been driven by evaluation of empirical performance. Theory may suggest that certain types of measurements and models may be effective, but it is empirical performance that matters. Why certain types of measurements and models work well may be difficult to understand from a theoretical perspective, but empirical testing demonstrates that they work well. Some phoneticians have criticized automatic approaches to forensic voice comparison, but empirical tests under forensically realistic conditions have demonstrated that the performance of systems based on automatic approaches is consistently much better than those based on acoustic-phonetic approaches. I do not believe that it would be helpful to raise this debate in the document, and in the recommended revisions I have removed the discussion of “processes”, and focused on black-box testing.

When forensic examiners do provide a statistical statement—with or without a numerical articulation of probability, odds, or likelihoods—such a statement must be supported by an empirical assessment of the underlying statistical model. Statistical calculations used in judicial proceedings should be replicable, given the data and statistical model; however, when observations are largely subjective or when different statistical models are in use, the quantitative summary of the significance of the findings may vary from examiner to examiner and from laboratory to laboratory. Consequently, an essential element of an examiner’s report is a statement of the measurements and the models to assist other experts in replicating the statistical quantities reported.

In comparing forensic evidence recovered from crime scenes or on victims or suspects with known samples, the forensic examiner primarily focuses on ascertaining corresponding features and, traditionally, in deciding whether there is a positive association (often referred to as a “match,” an “inclusion,” or “consistent” or “indistinguishable” features) or a negative one (an “exclusion” or inconsistency) to the known sample. But a “positive association” is not probative unless it is more probable when the items have a common source than when they originate from different sources. Indicating the statistical weight of the positive association therefore requires a statement of how common or rare the association is, based on a database linked to the case at hand. For example, a positive association for the presence or absence of pigment in a hair cuticle is some evidence that the hairs have a common origin, but the significance of this association is unknown without data from relevant populations.

More generally, when dealing with features, such as the refractive index of glass or the heights of peaks in an electropherogram of DNA fragments, that have more values than “absent” and “present,” the classification of “matching” and “not matching” omits statistical information related to the degree of similarity. The weight that should be given to any degree of association depends on (1) the probability of the degree of correspondence in the features, given that the samples came from the same source, and (2) the probability for the same measurement, given that the samples came from different sources. When the former probability is larger than the latter—when the observed degree of similarity occurs much more often for same-source samples than for different-source samples—the evidence supports the conclusion of a common source.

COMMENT:

“degree of association” is not a defined statistical term. I do not use it in the suggested revision.

I already incorporated a version of the penultimate sentence in an earlier paragraph.

I find the final sentence above potentially misleading in that it only discusses whether the value of a likelihood ratio is greater than or less than 1. A likelihood ratio is not a dichotomous indicator. The value of a likelihood ratio expresses the strength of the evidence, the further the value from 1 the greater the strength of the evidence. In the recommended revision I have removed this, but have not provided a replacement.

Any recommendation on presenting explicit probabilities, however derived for specific forensic evidence, might distinguish between probabilities based on some statistical model and ones that characterize the examiner’s subjective sense of how probable the evidence is under alternative hypotheses. The latter are difficult to validate, but it also must be understood that statistical models are approximations, and, inevitably, there is some uncertainty in the selection of a model. Furthermore, the statistical model and method used to analyze the evidence do not always admit naturally to the simple form of the likelihood ratio favored in the forensic science literature. Some statistical problems, especially those focused on issues of causation, may not involve source comparisons leading to likelihood ratios. In light of the limitations on both statistical modeling and more intuitive judgments of the significance of similarities, we offer the following views on the presentation of forensic science findings:

#### COMMENT:

I strongly disagree with the statement that “the statistical model and method used to analyze the evidence do not always admit naturally to the simple form of the likelihood ratio favored in the forensic science literature”. Introductory literature on the likelihood ratio framework may be simplified because it is introductory, a case in point is that the present document has focused on same-origin different-origin scenarios, but the framework is widely applicable and not restricted to source comparison. Statistical models used to calculate likelihood ratios may be anything but simple – state of the art models in, for example, DNA mixture analysis and forensic voice comparison deal with difficult problems and the models are concomitantly complex in order to effectively deal with those problems.

#### RECOMMENDED REVISION:

Which *statistical models* are appropriate for calculating a likelihood ratio as a statistical statement of the strength of evidence will depend on the type of data to be analyzed. Appropriate models for calculating a likelihood ratio in a forensic voice comparison case and in a DNA mixture case will be very different, because the type of measurements made on voice recordings are very different from DNA profiles (they have different data structures). As a matter of *transparency*, and to allow for *replication*, the forensic practitioner should document the statistical model which they employed.

If the data can only take on *discrete* values, then an appropriate model will calculate the relative *probabilities* of obtaining the observed values for the questioned-origin specimen if it came from the known source versus if it came from another source in the relevant population. If the data are *continuous* (they can take on any value), then an appropriate model will calculate the relative *likelihoods* of obtaining the observed values for the questioned-origin specimen if it came from the known source versus if it came from another source in the relevant population. In both cases, data from the known-source sample will be used to train a model to calculate the numerator of the likelihood ratio (the similarity term), and data from the relevant-population sample will be used to train a model to calculate the denominator of the likelihood ratio (the typicality term).

The value of each model (the probability or the likelihood estimate) will then be evaluated at the value of the datum from the questioned-source specimen. This results in two values, the numerator and the denominator for the likelihood ratio respectively. The numerator is divided by the denominator to arrive at a value for the likelihood ratio as a statistical statement quantifying the strength of evidence. [FN: For simplicity, we assumed that the questioned-source specimen provided a single data point. There are models for dealing with multiple data points from questioned-source specimens. Some models calculate the relative probabilities or likelihoods of obtaining the observed values for the questioned-source specimen and the known-source sample if they both had the same source versus if they had different sources (note that this is subtly different from what we described above). Some models do not directly calculate numerators and denominators as described above but still produce results which are interpretable as likelihood ratios answering the question specified by the two competing hypotheses.]

A traditional approach to evaluation of strength of evidence uses a “match” / “non-match” decision. If a “match” is declared, the strength of evidence can be evaluated using a model (trained on relevant data) that calculates an estimate of the relative probabilities of declaring a “match” when two objects really have the same source versus when they really have different sources (sensitivity divided by false-alarm rate). If the data are not intrinsically discrete, and the data are dichotomized into “match” and “non-match” by imposing a threshold on continuously valued data, such a model fails to exploit relevant information and is therefore suboptimal. More appropriate models which directly analyze continuously valued data should be used instead (see Morrison, Kaye, et al., 2016).

The likelihood ratio value which is calculated by the statistical model is an estimate of the strength of the evidence. Several factors will affect the performance of the system that makes those calculations, including how representative the data are of the relevant population, how good the chosen statistical model is at exploiting the information in the data, how much useful information there actually is in the measurements that are made, and how good or poor the conditions of the questioned-source specimen and known-source sample are. How well the system works should be assessed via *empirical validation*. The system includes the measurement procedures and the statistical models and any substantial actions taken by the forensic practitioner as part of the analysis. Only after the validity and reliability of the system have been empirically assessed and found to be adequate, can the inference as to the strength of evidence generated by the system be considered credible.

The *validity* and *reliability* (*accuracy* and *precision*) of the system which calculates the likelihood ratio should be empirically assessed under conditions reflecting those of the case under investigation. As with the data used for calculating the strength of evidence, the data for empirically testing the performance of the system should be sufficiently representative of the relevant population and sufficiently reflective of the questioned-source specimen and the known-source sample in the case under investigation that the results of testing will be meaningful with respect to how well the system is likely to perform in that case. The forensic practitioner should clearly communicate the nature of the test data so that the judge at an admissibility hearing and/or trier of fact at trial can decide if they were sufficiently representative of the relevant population and reflective of the conditions of the questioned-source specimen and the known-source sample in the case. The test data should consist of a large number of test pairs. One member of each test pair should reflect the conditions of the questioned-source specimen and the other those of the known-source sample. A number of these pairs must be known by the tester to be same-source pairs and the remainder must be known by the tester to be different-source pairs. The system being tested must not know the truth about each test pair. The tester presents each pair to the system, the system responds with a likelihood ratio, and the tester then assesses how good the response is given the tester’s knowledge about whether the pair was a same- or a different-source pair. For a same-source pair, the larger the likelihood ratio the better the performance, and for a different-source pair, the smaller the likelihood ratio the better the performance. The tester averages how good the performance is over all pairs. If the judge and/or trier of fact is satisfied that the test data were appropriate and that the number of test pairs was sufficient to provide a convincing assessment of performance, they can then consider whether the average performance of the system is good enough.

The previous paragraph described a procedure for assessing a measure of the validity or accuracy of the system. *Validity or accuracy* is a measure of *how close on average the answer is to the correct answer* (in this

context correct relates to whether the input was same or different source). *Reliability or precision* is a measure of *how consistent the answer is*. For example, if a different sample of the same relevant population or a different sample from the same known source were substituted, by how much would the value of the calculated likelihood ratio change? Reliability or precision quantifies the spread in the answers – two different systems could have the same accuracy (the same average performance), but one could be very precise (all the calculated likelihood ratio values cluster close together) and another relatively imprecise (the values have a wide spread). Assessing precision is somewhat more complicated than assessing accuracy, but can be done by using multiple samples from the same population and/or known source in the test data. [FN: The best way to deal with imprecision in forensic likelihood ratios is currently a matter of debate. Part of that debate appears in a 2016–2017 virtual special issue of the journal *Science & Justice* <http://www.sciencedirect.com/science/journal/13550306/vsi>]

In some branches of forensic science it may be possible to perform empirical validation under a set of conditions ahead of time, and then conduct analyses in a large number of cases which are sufficiently similar to the conditions under which the validation was performed. In other branches of forensic science the variability in relevant population and conditions may be so great from case to case that it may be necessary to conduct empirical validation on an essentially case-by-case basis. [FN: For more detailed coverage of empirical validation of forensic evaluation systems which produce likelihood ratios, see Morrison (2011) and Meuwly et al. (2016).]

Empirical validation treats the system to be tested as a *black box*. That is, it is concerned with how well the system works not with how the system works – it is not concerned with what goes on inside the box. In this sense it treats all systems equally.

As previously mentioned, for many types of evidence, forensic practitioners may not currently be making statistical assessments explicitly, but they may nevertheless be presenting their findings in a manner that connotes a statistical assessment. These assessments are based on practitioners' *subjective judgments* informed by their *training and experience*. Training and experience are important, but they are not a sufficient basis for establishing the validity and reliability of a practitioner's assessment of strength of evidence. There is nothing to prevent practitioners from assigning probabilities for the numerator of the likelihood ratio (the similarity term) and for the denominator of the likelihood ratio (the typicality term) on the basis of their subjective judgment. They should, however, be transparent as to the nature of their assignment of probabilities (the court may ask then to justify their choices), [FN: As part of transparency, and to demonstrate that the practitioner has actually considered both similarity and typicality with respect to the relevant population, we recommend that practitioners be required to report the values they assigned to both the numerator and the denominator of the likelihood ratio, not just the final value of the likelihood ratio itself.] and the performance of each forensic practitioner should be empirically validated under conditions reflecting those of the case under investigation. Each practitioner is treated as a system to be tested. Empirical testing is black box, it does not matter whether the system is based on quantitative measurements and statistical models or on a practitioner's subjective judgement. They are treated the same, and no system should be excused from the requirement to be empirically validated.

In light of the discussion above, we offer the following views on the presentation of forensic science findings:

### **Views of the Commission**

It is the view of the Commission that:

#### **COMMENT:**

Again, I make some between-paragraph comments, but then recommend a new set of views based on the recommended revisions above.

My recommended revision for this section is succinct. I think this section will be more effective if it is

succinct. I believe that the relevant details have already been covered in the overview section above.

In the recommended revision I deliberately use the term “should” rather than “must” since I think the subcommittee’s intention was to allow for some flexibility.

1. Forensic experts, both in their reports and in testimony, should present and describe the features of the questioned and known samples (the data), and similarities and differences in those features as well as the process used to arrive at determining them. The presentation should include statements of the limitations and uncertainties in the measurements or observations.

COMMENT:

As already discussed, similarities (or differences) are insufficient to quantify strength of evidence.

I interpret “describe the features” as a requirement to explain what measurements were made.

“limitations and uncertainties” I find to be vague terms. I think the concerns are addressed by empirical validation.

2. No one form of statistical calculation or statement is most appropriate to all forensic evidence comparisons or other inference tasks. Thus, the expert needs to be able to support, as part of a report and in testimony, the choice used in the specific analysis carried out and the assumptions on which it was based. When the statistical calculation relies on a specific database, the report should make clear which one and its relevance for the case at hand.

COMMENT:

I strongly disagree with the statement that “no one form of statistical statement is most appropriate”. A likelihood ratio is most appropriate.

That different models are appropriate for different data structures was already discussed above. I do not believe it is appropriate to repeat this in the views section – I think it a fact rather than a view.

3. The expert should report the limitations and uncertainty associated with measurements and the inferences that could be drawn from them. This report might take the form of an interval for an estimated value, or of separate statements regarding errors and uncertainties associated with the analysis of the evidence. If the expert has no information on sources of error in measurements and inferences, the expert must state this fact.

COMMENT:

Limitations and uncertainties, again, are incorporated in empirical validation.

I think that at this time it best not to prescribe how precision of systems calculation likelihood ratios (even if providing options) since, as discussed above, it is currently a matter of debate among forensic statistician and a consensus has not yet emerged. I think it is subsumed under empirical validation and statistical models for calculating likelihood ratios, and therefore does not need to be elaborated upon here.

“If the expert has no information ...” I think it best not to imply that empirical validation is optional.

4. Forensic science experts should not state that a specific individual or object is the source of the forensic science evidence and should make it clear that, even in circumstances involving extremely strong statistical evidence, it is possible that other individuals or objects could possess or have left a similar set of observed features.<sup>4</sup> Forensic science experts should confine their evaluative statements to the support that the findings provide for the claim linked to the forensic evidence.

---

<sup>4</sup>Similarly, to avoid implying that a statistical foundation exists when there is no statistical model or method and related database to properly characterize the evidence, forensic experts should not use phrases such as “to a reasonable degree of scientific certainty.” Nat’l Comm’n on Forensic Science, Views Document on Use of the Term “Reasonable Scientific Certainty,” Mar. 22, 2016, <https://www.justice.gov/ncfs/file/839731/download>.

#### COMMENTS:

I have put a “don’t” item following the list of “do” items.

Since there is already a whole NCFS document prohibiting the use of the phrase “reasonable scientific certainty”, I question whether it is necessary to repeat it here – it might actually be distracting. There are also many other inappropriate phrases used by practitioners (see Jackson, 2009).

5. To explain the value of the data in addressing claims as to the source of a questioned sample, forensic examiners may:
  - A. Refer to relative frequencies of individual features in a sample of individuals or objects in a relevant population (as sampled and then represented in a reference database). The examiner should note the uncertainties in these frequencies as estimates of the frequencies of particular features in the population.

#### COMMENTS:

Relative frequencies may be appropriate for estimating probabilities for discrete data, but for continuously valued data probability density models are needed to estimate likelihoods.

- B. Present estimates of the relative frequency of an observed combination of features in a relevant population based on a probabilistic model that is well grounded in theory and data. The model may relate the probability of the combination to the probabilities of individual features.
    - C. Present probabilities (or ratios of probabilities) of the observed features under different claims as to the origin of the questioned sample. The examiner should note the uncertainties in any such values.
    - D. When the statistical statement is derived from an automated computer-based system for making classifications, present not only the classification but also the operating characteristics of the system (the sensitivity and specificity of the system as established in

relevant experiments using data from a relevant population). If the expert has no information or limited information about such operating characteristics, the expert must state this fact.

#### COMMENTS:

Systems that make classifications should not be used – this is a version of the “match”/“non-match” approach. “If the expert has no information ...” Again, I think it best not to imply that empirical validation is optional.

6. Not all forensic subdisciplines currently can support a probabilistic or statistical statement. There may still be value to the factfinder in learning whatever comparisons the expert in those subdisciplines has carried out. But the absence of models and empirical evidence needs to be expressed both in testimony and written reports.

#### RECOMMENDED REVISION:

##### Views of the Commission

It is the view of the Commission that:

Forensic practitioners should evaluate strength of evidence using relevant data and statistical models. In both their reports and testimony, they should:

1. Clearly communicate the two competing hypotheses they set out to evaluate, including (whenever applicable) the relevant population specified as part of one of the hypotheses.
2. Describe the conditions of the samples and/or specimens (or other form of data) they were asked to analyze.
3. Explain how they obtained sample data, for model training and for system testing, that are (as applicable) representative of the relevant population and reflective of the conditions of the samples and/or specimens they were asked to analyze.
4. Describe (if the raw data were not numeric) the quantitative measurement procedures they used in order to generate numeric data.
5. Describe the statistical models they used to calculate likelihood ratios from the numeric data.
6. Describe the procedures and data they used to empirically validate system performance, and present the results of that empirical validation.
7. Present a likelihood ratio as a statistical statement which quantitatively expresses the strength of evidence associated with the samples and/or specimens (or other data) they were asked to analyze.

Forensic practitioners’ statements of strength of evidence should be restricted to statements which are logically correct and justified via inference from relevant data. For example, forensic practitioners should not state or imply that a specific individual or object *is* the source of a questioned specimen, they should not make statements that refer only to similarity and not also to typicality, and they should not use expressions such as “to a reasonable degree of scientific certainty”. [FN: Nat’l Comm’n on Forensic Science, Views Document on Use of the Term “Reasonable Scientific Certainty,” Mar. 22, 2016, <https://www.justice.gov/ncfs/file/839731/download>]

If, rather than using a statistical model, a forensic practitioner evaluates strength of evidence using their subjective judgment informed by their training and experience, then in place of step 5 above they should (1) state that subjective judgment is the basis for their assignment of probabilities, and (2) state the probability values they assigned to each of the numerator and the denominator of the likelihood ratio and their reasons

for assigning those particular values. They should otherwise conform to all the steps above, including the requirement for empirical validation.

#### REFERENCES:

- Jackson G. (2009). Understanding forensic science opinions. Ch. 16 (pp. 419–445) in Fraser J., Williams R. (Eds.), *Handbook of Forensic Science*. Cullompton, UK: Willan.
- Meuwly D., Ramos D., Haraksim D. (2016). A guideline for the validation of likelihood ratio methods used for forensic evidence evaluation, *Forensic Science International*.  
<http://dx.doi.org/10.1016/j.forsciint.2016.03.048>
- Morrison G.S. (2011). Measuring the validity and reliability of forensic likelihood-ratio systems. *Science & Justice*, 51, 91–98. <http://dx.doi.org/10.1016/j.scijus.2011.03.002>
- Morrison G.S., Kaye D.H., Balding D.J., Taylor D., Dawid P., Aitken C.G.G., Gittelsohn S., Zadora G., Robertson B., Willis S., Pope S., Neil M., Martire K.A., Hepler A., Gill R.D., Jamieson A., de Zoete J., Ostrum R.B., Caliebe A. (2016). A comment on the PCAST report: Skip the “match”/“non-match” stage. *Forensic Science International*. <http://dx.doi.org/10.1016/j.forsciint.2016.10.018>
- Morrison G.S., Thompson W.C. (2017). Assessing the admissibility of a new generation of forensic voice comparison testimony. *Columbia Science and Technology Law Review*, 18. Preprints available at <https://ssrn.com/abstract=2883767> and <https://www.newton.ac.uk/files/preprints/ni16053.pdf>
- Robertson B., Vignaux G.A., Berger C.E.H. (2016). *Interpreting Evidence: Evaluating Forensic Science in the Courtroom* (2nd ed.). Chichester, UK: Wiley.

#### COMPLETE REDRAFT INCLUDING ALL RECOMMENDED REVISIONS:

begins next page

## Overview

This Views document presents background information and views on the following question: How should forensic practitioners evaluate and report *strength of evidence*?<sup>1</sup> The statements of strength of evidence that forensic practitioners make in reports and testimony are referred to in this document as *statistical statements*. Such statements are made in relation to many types of forensic evidence. Five examples follow.

1. *Pattern and impression evidence*. A shoe print found at a crime scene is compared with the sole of the defendant's shoe. How strongly do the results of the examination support the prosecution's contention that the defendant's shoe is the source of the crime-scene print versus the defense's contention that it is not?

2. *Trace evidence*. A burglar smashed a window. Various physical and chemical properties of fragments of glass collected from the suspect's clothing and fragments from the broken window at the crime scene are measured and compared. How strongly do the results of the analysis support the prosecution's contention that the broken window at the crime scene is the source of the fragments found on the suspect's clothing as opposed to the suspect's claim that the fragments are from a different source?

3. *Activity level*. A burglar smashes a window. Four hours later a suspect is arrested. Two fragments of glass are found in the suspect's jacket pocket. The suspect denies involvement in the burglary, but says he did accidentally break a window in the same building two weeks earlier (he works as a window cleaner). How probable is it that the suspect would have two fragments of glass in his pocket if he had broken a window four hours earlier and two weeks earlier versus if he had only broken a window two weeks earlier?

4. *Quantitative analysis*. A chemical analysis of fire debris is conducted in an arson investigation. Results are found that are thought to be positive indicators for the presence of ignitable liquid residue. What is the probability of getting these results if ignitable liquid really had been present before the fire started versus if ignitable liquid had not been present (positive results have been known to occur for other reasons).

5. *Quantitative analysis with extrapolation*. A blood sample is collected from a driver 2 hours after an accident. A chemical test indicates a blood alcohol concentration (BAC) of 0.04%, which is below the 0.08% legal limit. How probable is it that the BAC at the time of the accident was above the legal limit? How probable is it that the BAC at the time of the accident was below the legal limit?

These statistical statements and those that appear in connection with many other forms of evidence should be based on *relevant data* and *appropriate statistical models*.

Statistics is concerned with the study of variability, probability, and decision-making in the face of uncertainty. It supplies a set of principles, based on logic, for drawing inferences from data.

For many types of evidence, forensic practitioners may not currently be making statistical assessments explicitly, but they may nevertheless be presenting their findings in a manner that connotes a statistical assessment. For example, the statement that "the latent print comes from the defendant's thumb" or "it is unlikely that the print came from anyone else" suggests a high probability that the print came from the defendant as determined by an understanding of the frequencies of similar features in fingerprints from the same individual and in prints from different individuals. As forensic science moves forward, the Commission anticipates efforts to make the presentation of analyses more overtly statistical and quantitative, and it is toward this end that the Commission expresses its views. The Commission advances a preference for an explicit statistical foundation for statements because a statistical analysis tends to provide a ready means for assessing and expressing uncertainty.

The framework we describe here applies throughout forensic science (including all of the examples listed above and more), but for simplicity and concreteness we focus on comparison of a specimen of questioned origin and a sample of known origin, e.g., a shoe print at a crime scene compared with a defendant's shoe, or fragments of glass recovered from a suspect's clothing compared with fragments of glass from a broken window at a crime scene (see examples 1 and 2 above). In such scenarios the question to be answered has two

---

<sup>1</sup> "strength of evidence" has also been referred to as "weight of evidence" and as "probative value"

parts:

1. What is the probability of the observed properties of the questioned-source specimen if it came from the known source?
2. What is the probability of the observed properties of the questioned-source specimen if it came not from the known source but from some other source in the relevant population?

Part 1 can be considered a question about how *similar* the questioned-source specimen is to the known source, and part 2 can be considered a question about how *typical* the questioned-source specimen is with respect to the relevant population.<sup>2</sup> Dividing the answer to part 1 by the answer to part 2 produces a *likelihood ratio*, which is a statistical statement of the strength of evidence.<sup>3</sup>

Note that just knowing about similarity is not particularly informative as to strength of evidence. For example, if the shoe print from the crime scene and the defendant's shoe are both the same size, they are maximally similar on this metric. If the relevant population were shoes worn by adult males in the United States, however, the strength of evidence would be relatively weak if they were both size 10, a very typical shoe size, but relatively strong if they were both size 14, a quite atypical shoe size.

*Relevant numeric data* will consist of *quantitative measurements* made on (a) the questioned-source specimen, (b) a sample of the known source, and (c) a sample of the relevant population.

The *relevant population* is the population from which the questioned-source specimen could have come if it had not come from the known source. For example, if the questioned-source specimen is a footwear print in the snow in Alaska in winter, the relevant population would at least be restricted to footwear worn in Alaska when there is snow on the ground in winter. Footwear worn in Hawai'i at any time of year would obviously not be the relevant population, but footwear worn in Alaska in summer would also not be the relevant population. Footwear worn in the city may not be the relevant population if the crime were committed in a rural area, etc., etc. If the questioned-source specimen is fragments of glass found on the suspect's clothing, and the suspect claims to have innocently broken another window in the same building as the crime scene, then ideally it would be glass from the particular other window that would be the relevant population. If the suspect makes no specific claims as to where the glass fragments may have come from, then the relevant population could be quite wide, including window glass in that part of the world and also automotive and container glass. What constitutes the relevant population in a given case may require substantial consideration.

Since the calculated strength of evidence will depend in part on typicality with respect to the relevant population, it is important that the forensic practitioner clearly communicate what they have specified as the relevant population. The judge at an admissibility hearing and/or the trier of fact a trial needs to know what relevant population the forensic practitioner specified in the typicality part of the question, first so that they can decide whether the forensic practitioner is actually answering a relevant question for them, and second so that they can understand the answer that the forensic practitioner provides to that question – if one does not understand the question, one cannot understand the answer.

In addition to specifying the relevant population they intend to use, the forensic practitioner must use a *sample* which is sufficiently representative of the relevant population. A sample which is too small or which is systematically biased will lead to a poor estimate of the probability of the observed properties of the questioned-source specimen if it came not from the known source but from some other source in the relevant population. This in turn will lead to a poor estimate of the likelihood ratio as a statistical statement of the strength of evidence. It may be impossible to obtain a sample that is perfectly representative of exactly the

---

<sup>2</sup> More generally, parts 1 and 2 are often referred to as the *prosecution hypothesis* and the *defense hypothesis* (or the *alternative hypothesis*) respectively. The term “defense hypothesis” does not imply that the defense must advance an alternative hypothesis, but in calculating strength of evidence the forensic practitioner must consider an alternative hypothesis that is some meaningful negation of the prosecution hypothesis. The two hypotheses must be mutually exclusive and, within reason, exhaustive.

<sup>3</sup> A substantial body of literature exists addressing theoretical aspects of the likelihood ratio framework and describing and evaluating statistical models for calculating likelihood ratios as quantifications of strength of evidence. A new edition of a classic 1995 book-length introduction to the likelihood ratio framework is: Robertson B., Vignaux G.A., Berger C.E.H. (2016). *Interpreting Evidence: Evaluating Forensic Science in the Courtroom* (2nd ed.). Chichester, UK: Wiley.

right population. The issue is whether the sample is sufficiently representative that the results of the analysis will be reasonable answers to the question asked. For example, in the broken window example, if glass fragments from the specific window that the suspect claims to have broken two weeks earlier are not available, samples of glass from other windows in the same building might be a reasonable practical alternative that implies a different but still meaningful question. Again, it is important that the forensic practitioner clearly communicate what they did so that their decisions and actions can be reviewed by the court.

In many branches of forensic science the *conditions* of the questioned specimen and known sample are relevant. For example, in a forensic voice comparison case the questioned-speaker recording may be an intercept of a lively mobile telephone conversation on which there is background traffic noise, and the known-speaker recording may be of subdued answers to police interview questions recorded in a small echoey room. Calculating a likelihood ratio based on data that reflect conditions substantially different from those of the questioned specimen and known sample may produce misleading results, hence the forensic practitioner will need to use data which are sufficiently reflective of the conditions of the questioned specimen and known sample in the case. Again, it is important that the forensic practitioner clearly communicate what they did so that their decisions and actions can be reviewed by the court.

Which *statistical models* are appropriate for calculating a likelihood ratio as a statistical statement of the strength of evidence will depend on the type of data to be analyzed. Appropriate models for calculating a likelihood ratio in a forensic voice comparison case and in a DNA mixture case will be very different, because the type of measurements made on voice recordings are very different from DNA profiles (they have different data structures). As a matter of *transparency*, and to allow for *replication*, the forensic practitioner should document the statistical model which they employed.

If the data can only take on *discrete* values, then an appropriate model will calculate the relative *probabilities* of obtaining the observed values for the questioned-origin specimen if it came from the known source versus if it came from another source in the relevant population. If the data are *continuous* (they can take on any value), then an appropriate model will calculate the relative *likelihoods* of obtaining the observed values for the questioned-origin specimen if it came from the known source versus if it came from another source in the relevant population. In both cases, data from the known-source sample will be used to train a model to calculate the numerator of the likelihood ratio (the similarity term), and data from the relevant-population sample will be used to train a model to calculate the denominator of the likelihood ratio (the typicality term). The value of each model (the probability or the likelihood estimate) will then be evaluated at the value of the datum from the questioned-source specimen. This results in two values, the numerator and the denominator for the likelihood ratio respectively. The numerator is divided by the denominator to arrive at a value for the likelihood ratio as a statistical statement quantifying the strength of evidence.<sup>4</sup>

A traditional approach to evaluation of strength of evidence uses a “match” / “non-match” decision. If a “match” is declared, the strength of evidence can be evaluated using a model (trained on relevant data) that calculates an estimate of the relative probabilities of declaring a “match” when two objects really have the same source versus when the really have different sources (sensitivity divided by false-alarm rate). If the data are not intrinsically discrete, and the data are dichotomized into “match” and “non-match” by imposing a threshold on continuously valued data, such a model fails to exploit relevant information and is therefore suboptimal. More appropriate models which directly analyze continuously values data should be used instead.<sup>5</sup>

The likelihood ratio value which is calculated by the statistical model is an estimate of the strength of the

---

<sup>4</sup> For simplicity, we assumed that the questioned-source specimen provided a single data point. There are models for dealing with multiple data points from questioned-source specimens. Some models calculate the relative probabilities or likelihoods of obtaining the observed values for the questioned-source specimen and the known-source sample if they both had the same source versus if they had different sources (note that this is subtly different from what we described above). Some models do not directly calculate numerators and denominators as described above but still produce results which are interpretable as likelihood ratios answering the question specified by the two competing hypotheses.

<sup>5</sup> see: Morrison G.S., Kaye D.H., Balding D.J., Taylor D., Dawid P., Aitken C.G.G., Gittelson S., Zadora G., Robertson B., Willis S., Pope S., Neil M., Martire K.A., Hepler A., Gill R.D., Jamieson A., de Zoete J., Ostrum R.B., Caliebe A. (2016). A comment on the PCAST report: Skip the “match”/“non-match” stage. *Forensic Science International*. <http://dx.doi.org/10.1016/j.forsciint.2016.10.018>

evidence. Several factors will affect the performance of the system that makes those calculations, including how representative the data are of the relevant population, how good the chosen statistical model is at exploiting the information in the data, how much useful information there actually is in the measurements that are made, and how good or poor the conditions of the questioned-source specimen and known-source sample are. How well the system works should be assessed via *empirical validation*. The system includes the measurement procedures and the statistical models and any substantial actions taken by the forensic practitioner as part of the analysis. Only after the validity and reliability of the system have been empirically assessed and found to be adequate, can the inference as to the strength of evidence generated by the system be considered credible.

The *validity* and *reliability* (*accuracy* and *precision*) of the system which calculates the likelihood ratio should be empirically assessed under conditions reflecting those of the case under investigation. As with the data used for calculating the strength of evidence, the data for empirically testing the performance of the system should be sufficiently representative of the relevant population and sufficiently reflective of the questioned-source specimen and the known-source sample in the case under investigation that the results of testing will be meaningful with respect to how well the system is likely to perform in that case. The forensic practitioner should clearly communicate the nature of the test data so that the judge at an admissibility hearing and/or trier of fact at trial can decide if they were sufficiently representative of the relevant population and reflective of the conditions of the questioned-source specimen and the known-source sample in the case. The test data should consist of a large number of test pairs. One member of each test pair should reflect the conditions of the questioned-source specimen and the other those of the known-source sample. A number of these pairs must be known by the tester to be same-source pairs and the remainder must be known by the tester to be different-source pairs. The system being tested must not know the truth about each test pair. The tester presents each pair to the system, the system responds with a likelihood ratio, and the tester then assesses how good the response is given the tester's knowledge about whether the pair was a same- or a different-source pair. For a same-source pair, the larger the likelihood ratio the better the performance, and for a different-source pair, the smaller the likelihood ratio the better the performance. The tester averages how good the performance is over all pairs. If the judge and/or trier of fact is satisfied that the test data were appropriate and that the number of test pairs was sufficient to provide a convincing assessment of performance, they can then consider whether the average performance of the system is good enough.

The previous paragraph described a procedure for assessing a measure of the validity or accuracy of the system. *Validity or accuracy* is a measure of *how close on average the answer is to the correct answer* (in this context correct relates to whether the input was same or different source). *Reliability or precision* is a measure of *how consistent the answer is*. For example, if a different sample of the same relevant population or a different sample from the same known source were substituted, by how much would the value of the calculated likelihood ratio change? Reliability or precision quantifies the spread in the answers – two different systems could have the same accuracy (the same average performance), but one could be very precise (all the calculated likelihood ratio values cluster close together) and another relatively imprecise (the values have a wide spread). Assessing precision is somewhat more complicated than assessing accuracy, but can be done by using multiple samples from the same population and/or known source in the test data.<sup>6</sup>

In some branches of forensic science it may be possible to perform empirical validation under a set of conditions ahead of time, and then conduct analyses in a large number of cases which are sufficiently similar to the conditions under which the validation was performed. In other branches of forensic science the variability in relevant population and conditions may be so great from case to case that it may be necessary to conduct empirical validation on an essentially case-by-case basis.<sup>7</sup>

---

<sup>6</sup> The best way to deal with imprecision in forensic likelihood ratios is currently a matter of debate. Part of that debate appears in a 2016–2017 virtual special issue of the journal *Science & Justice* <http://www.sciencedirect.com/science/journal/13550306/vsi>

<sup>7</sup> For more detailed coverage of empirical validation of forensic evaluation systems which produce likelihood ratios, see: Morrison G.S. (2011). Measuring the validity and reliability of forensic likelihood-ratio systems. *Science & Justice*, 51, 91–98. <http://dx.doi.org/10.1016/j.scijus.2011.03.002>. Meuwly D., Ramos D., Haraksim D. (2016). A guideline for the validation of likelihood ratio methods used for forensic evidence evaluation, *Forensic Science International*.

Empirical validation treats the system to be tested as a *black box*. That is, it is concerned with how well the system works not with how the system works – it is not concerned with what goes on inside the box. In this sense it treats all systems equally.

As previously mentioned, for many types of evidence, forensic practitioners may not currently be making statistical assessments explicitly, but they may nevertheless be presenting their findings in a manner that connotes a statistical assessment. These assessments are based on practitioners' *subjective judgments* informed by their *training and experience*. Training and experience are important, but they are not a sufficient basis for establishing the validity and reliability of a practitioner's assessment of strength of evidence. There is nothing to prevent practitioners from assigning probabilities for the numerator of the likelihood ratio (the similarity term) and for the denominator of the likelihood ratio (the typicality term) on the basis of their subjective judgment. They should, however, be transparent as to the nature of their assignment of probabilities (the court may ask then to justify their choices),<sup>8</sup> and the performance of each forensic practitioner should be empirically validated under conditions reflecting those of the case under investigation. Each practitioner is treated as a system to be tested. Empirical testing is black box, it does not matter whether the system is based on quantitative measurements and statistical models or on a practitioner's subjective judgement. They are treated the same, and no system should be excused from the requirement to be empirically validated.

In light of the discussion above, we offer the following views on the presentation of forensic science findings:

### **Views of the Commission**

It is the view of the Commission that:

Forensic practitioners should evaluate strength of evidence using relevant data and statistical models. In both their reports and testimony, they should:

1. Clearly communicate the two competing hypotheses they set out to evaluate, including (whenever applicable) the relevant population specified as part of one of the hypotheses.
2. Describe the conditions of the samples and/or specimens (or other form of data) they were asked to analyze.
3. Explain how they obtained sample data, for model training and for system testing, that are (as applicable) representative of the relevant population and reflective of the conditions of the samples and/or specimens they were asked to analyze.
4. Describe (if the raw data were not numeric) the quantitative measurement procedures they used in order to generate numeric data.
5. Describe the statistical models they used to calculate likelihood ratios from the numeric data.
6. Describe the procedures and data they used to empirically validate system performance, and present the results of that empirical validation.
7. Present a likelihood ratio as a statistical statement which quantitatively expresses the strength of evidence associated with the samples and/or specimens (or other data) they were asked to analyze.

Forensic practitioners' statements of strength of evidence should be restricted to statements which are logically correct and justified via inference from relevant data. For example, forensic practitioners should not state or imply that a specific individual or object *is* the source of a questioned specimen, they should not make statements that refer only to similarity and not also to typicality, and they should not use expressions such as "to a reasonable degree of scientific certainty".<sup>9</sup>

---

<http://dx.doi.org/10.1016/j.forsciint.2016.03.048>

<sup>8</sup> As part of transparency, and to demonstrate that the practitioner has actually considered both similarity and typicality with respect to the relevant population, we recommend that practitioners be required to report the values they assigned to both the numerator and the denominator of the likelihood ratio, not just the final value of the likelihood ratio itself.

<sup>9</sup> Nat'l Comm'n on Forensic Science, Views Document on Use of the Term "Reasonable Scientific Certainty," Mar. 22, 2016, <https://www.justice.gov/ncfs/file/839731/download>

If, rather than using a statistical model, a forensic practitioner evaluates strength of evidence using their subjective judgment informed by their training and experience, then in place of step 5 above they should (1) state that subjective judgment is the basis for their assignment of probabilities, and (2) state the probability values they assigned to each of the numerator and the denominator of the likelihood ratio and their reasons for assigning those particular values. They should otherwise conform to all the steps above, including the requirement for empirical validation.