Raymond Marquis, Alex Biedermann, Liv Cadola, Christophe Champod, Line Gueissaz, Genevi`eve Massonnet, Williams David Mazzella, Franco Ta- roni, Tacha Hicks, Discussion on how to implement a verbal scale in a forensic laboratory: Benefits, pitfalls and suggestions to avoid misunderstandings, Science & Justice, 56, 364-370, doi: http://dx.doi.org/10.1016/j.scijus.2016.05.009.

Title

Discussion on how to implement a verbal scale in a forensic laboratory: Benefits, pitfalls and suggestions to avoid misunderstandings

Abstract

In a recently published guideline for evaluative reporting in forensic science, the European Network of Forensic Science Institutes (ENFSI) recommended the use of the likelihood ratio for the measurement of the value of forensic results. As a device to communicate the probative value of the results, the ENFSI guideline mentions the possibility to define and use a verbal scale, which should be unified within a forensic institution. This paper summarizes discussions held between scientists of our institution to develop and implement such a verbal scale. It intends to contribute to general discussions likely to be faced by any forensic institution that engages in continuous monitoring and improving of their evaluation and reporting format. We first present published arguments in favour of the use of such verbal qualifiers. We emphasize that verbal qualifiers do not replace the use of numbers to evaluate forensic findings, but are useful to communicate the probative value, since the weight of evidence in terms of likelihood ratio are still apprehended with difficulty by both the forensic scientists, especially in absence of hard data, and the recipient of information. We further present arguments that support the development of the verbal scale we propose. Recognising the limits of the use of such a verbal scale, we then discuss its disadvantages: it may lead to the spurious view according to which the value of the observations made in a given case is relative to other cases. Verbal qualifiers are also prone to misunderstandings and cannot be coherently combined with other evidence. We therefore recommend not using the verbal qualifier alone in a written statement. While scientists should only report on the probability of the findings – and not on the probability of the propositions, which are the duty of the Court – we suggest showing examples to let the recipient of information understand how the scientific evidence affects the probabilities of the propositions. To avoid misunderstandings, we also advise to mention in the statement what the results do not mean. Finally, we are of the opinion that if experts were able to coherently articulate numbers, and if recipients of information could properly handle such numbers, then verbal qualifiers could be abandoned completely. At that time, numerical expressions of probative value will be appropriately understood, as other numerical measures that most of us understand without the need of any further explanation, such as expressions for length or temperature.

Keywords

Verbal scale, Interpretation, Likelihood ratio, Probative value, Statement

1. Introduction

In a recently published guideline for evaluative reporting[1] in forensic science, the European Network of Forensic Science Institutes [2] recommended the use of the likelihood ratio for the measurement of the value of forensic results[2]. The document specifies a series of principles to guide the scientist's thinking about the evaluation of forensic results: the first principle is that interpretation takes place in a framework of circumstances and that the value of forensic observations depends on relevant case information. In other words, if the case information changes, the interpretation of the findings must be reviewed as well. The second principle states that the forensic observations shall be interpreted in the light of at least one pair of competing propositions. The third principle stresses that it is appropriate for the scientist only to address the probability of the observations given the propositions, and not the probability of the propositions themselves [3,4,5].

The application of these principles ensures that the approach is balanced and logical. In particular, with the value of the likelihood ratio, scientists express an opinion on the observations they have made and they convey the degree of support provided by these results for one proposition over the alternative. The recipients of expert information (e.g., the Court) can then use this information to update their belief on the competing propositions, considering all the other elements of the case.

Although the approach is well structured, its practical implementation represents a challenge to both individual scientists and forensic science institutions as a whole. The introduction of such a change in evaluation and reporting practice does not happen overnight and requires an institutional strategy over a period of time[3].

As a device to support scientists' reporting practice, the ENFSI guideline mentions the possibility to define and use standardised verbal qualifiers for ranges of likelihood ratio values — also often referred to as a 'verbal scale'. There is no binding recommendation in the ENFSI guideline, but it is advised to use a single and unified reporting convention for all disciplines within a forensic institution. It must be emphasised that a verbal scale is not a replacement for the likelihood ratio value, but it can represent a convenient way to communicate this value.

This article reports on this particular aspect of forensic science reporting. Specifically, we will focus on questions and discussions that emerged from works towards the development and implementation of optional verbal qualifiers for probative value. This paper — organised in a question-answer format — intends to contribute to general discussions likely to be faced by any forensic institution that engages in continuous monitoring and improving of their evaluation and reporting format. We hope that our readers will see merit in this initiative, as we believe that sharing practical experience from different institutions regarding challenges, approaches and strategies for implementing the principles emphasized in the ENFSI guideline is essential.

---

[1] Please note that our discussions will neither include investigative nor technical reporting as defined in both the statement of the Association of Forensic Science Providers [1] and the recent ENFSI Guideline [2].

[2] A likelihood ratio is a ratio of two probabilities: the probability of the observations given that the first proposition and the conditioning information are true, divided by the probability of the observations given that the alternative proposition and the conditioning information are true.

[3] For this purpose, the ENFSI guideline proposes a general four step roadmap to help quality managers and leading scientists design an implementation plan that is flexible enough to be adapted to service specific requirements and needs.

2. Discussion of key questions

In the following sections, we describe the key elements raised in the discussions held within our institution. As it appears that verbal scales to communicate the weight of evidence are quite commonly used in the forensic community, we highlight in section 2.1 the arguments in favour of such a practice. We then address, in section 2.2, the question of whether such a verbal scale remains useful in the current state of practice. As it emerges that a verbal scale may still be of help today, section 2.3 discusses the question of what type of verbal scale should be used. Section 2.4 covers misunderstandings that occur when expert opinion is only conveyed by words. To overcome these problems, one possible solution could be to add the full range of the verbal scale in the statement. Section 2.5 provides arguments against this suggestion. Finally, section 2.6 proposes some recommendations that can help communicating the value of evidence.

2.1. What are the published arguments in favour of the use of a verbal scale as suggested in the ENFSI guideline for evaluative reporting in forensic science?

Harmonisation of conclusions and the use of common terminology is an important aspect in forensic science reporting. As early as 1979, Kind *et al.* [6] suggested conventions regarding categorisation of "*samples*" submitted to forensic science laboratories for examination. In 1986, Brown and Cropp [7] discussed the importance of avoiding some terms in reports and even went a step further by suggesting harmonisation of conclusions using a correspondence table between probabilities and adverbs. A year later, Leung and Cheung [8] commented on the wide range of terminologies used to define qualified opinions, and proposed a way to provide uniformity in this respect. One can see therefore that standardisation of qualitative terms to report the opinion of forensic scientists regarding the value of evidence has been a preoccupation for forensic scientists for a long time.

Authors focused on the choice of terms used to conclude and tried to find appropriate words to convey the value of the evidence in a case, but to our knowledge it was Evett [9] who first suggested adopting a verbal scale in forensic science in 1987, based on Jeffreys's book 'Theory of Probability', first published in 1939. We summarize hereafter arguments presented in favour of the use of a verbal scale.

Early arguments in favour of the use of a so-called 'verbal scale' are that it should promote logical reporting. Indeed, according to Evett [9], verbal qualifiers can help scientists express themselves on the value of the results given the propositions rather than on propositions themselves. By saying "the results [strongly] support defence proposition rather than prosecution proposition", scientists ensure that they do not transpose the conditional. The forensic observations are thus evaluated in agreement with a logical approach, and the scientist's statement of the value of the observations serves the purpose to assist the recipient of expert information going from prior to posterior odds [3].

A further argument in favour of the use of verbal equivalents is that in fields where structured, documented and published data are scarce, reporting a likelihood ratio may be a challenge. Indeed, some scientists feel that they can only commit themselves to particular numbers if they can trace them back to calculations based on hard numerical data [10,11]. Also it may give the illusion of a level of mathematical precision, whereas no calculation was carried out per se, but only an assignment based on experience and training. Instead, in those situations, scientists are generally more at ease by communicating the magnitude of their likelihood ratio with a verbal equivalent. For example, Jackson [10] wrote that "the scientist

should evaluate broadly the magnitude for the likelihood ratio and translate that into a verbal equivalent" (p.85).

Verbal conventions also cover an important third function, as one assumes that it helps communication between scientists and non-scientists. Indeed, faced with uncertainty, it is known that generally people prefer words to numbers [12] so that the use of verbal equivalents is considered to be beneficial for the recipients of expert information who "not feel confident in handling numbers and react negatively to mathematical formulae" (p.447) [13].

For the above reasons, it would seem convenient for scientists to use a verbal equivalent in their statement, instead of reporting the (numerical) value of the likelihood ratio alone: this verbal term would then express the value of the observations as given by the likelihood ratio. But, as we will see in the next sections, such a device for supporting reporting has some disadvantages that one needs to consider.

2.2. Do we need verbal equivalents in the current state of practice?

One negative aspect of verbal qualifiers is that scientists may fail to remember that "the assessment comes first, the decision about a verbal qualifier comes later" (p.47) [14]. Ideally, as discussed by Berger *et al.* [14], the assignment of a numerical value, that is a probability (or probability density), is "preferable whenever possible" (p.47). It is true that likelihood ratios can be assigned qualitatively, but experts need to express themselves in numerical terms as it is through numbers that we measure things, convey information and combine it with other information [12]. Without numbers, one cannot combine different types of traces if needed. One cannot really appreciate either the impact of the additional information on how the recipient of information evaluates both propositions. Moreover, numbers allow us to make distinctions that words cannot make: indeed, the range of practical situations that ask us to make and convey distinctions is so large that we virtually run out of words especially with very high likelihood ratios. So, only numbers can cope with this challenge. Numbers also offer the advantage of ensuring that when different experts use the same agreed term, they intend to convey the same message. The choice can be left to the expert to communicate a verbal equivalent instead of a number alone (or an order of magnitude of this number), but it would be wrong to say that numbers are not essential to this approach. This misconceives the intricacy of practical needs and the measurement problem in the first place (i.e., the measurement of uncertainty).

In some areas of forensic science, sufficient data and agreed models are available for actually *calculating* or *computing* a likelihood ratio. This is the case for DNA profiling, for example [15]. Berger *et al.* [14] also contend that "[i]n those cases where a quantitative likelihood ratio has been calculated, (…) it is the number alone that should be put to the jury" (p.47).

Probability elicitation when there are no data but essentially expert knowledge (for example in the disciplines involving pattern analysis) can be more difficult. There is much effort still to be made in training of experts in the assignment of probabilities in casework. As Stoney [16] put it more than 30 years ago, "[e]ven if there is general agreement that the likelihood ratio is the appropriate method to evaluate physical evidence, this does not solve the question of which probabilities are appropriate to use in the ratio itself" (p. 480). However, even if these probabilities can legitimately be informed by experience, "it will be necessary for the practitioner to be able to demonstrate which experiences provide basis for those probabilities" (p. 213) [17]. Assigning a number to express the value of the observations has the advantage to force scientists to elicit their thinking process. They will necessarily spend more time to interpret their findings and will consequently be more prepared for court hearings. It goes without saying that the basis for such a number (or its order of magnitude)

shall be explained in the written statement so to be fully transparent, as one would if only giving a verbal term.

Suppose that an expert is able to coherently articulate a number, even if this is based on documented knowledge and experience rather than on 'hard' data. Suppose also that communication was improved to the point that recipients of expert information were able to properly use the number conveyed by the expert within the framework of a logical approach. Would we then still need verbal qualifiers? From a logical point of view, the answer is no. To illustrate this, think of widely known systems of measurement such as weight and length. Most people are acquainted with expressions such as '1 kilogram' of something (e.g., rice) or '1 meter' (e.g., for the length of a table). We understand such quantifications because we are literate with the measured concepts and the units of measurement. The same could be expected if people were perfectly acquainted with matters concerning the measurement of uncertainty using probability. Indeed, it is rather surprising to observe the limited capacity of people for such an important concept (i.e., probability) to capture an essential feature of life (i.e., uncertainty). By extension, we should also be able to expect that the notion of likelihood ratio as a measure of probative strength would benefit from such improved understanding.

However, as long as this is not the case, it appears useful or maybe simply reassuring to maintain verbal qualifiers to help - along with other strategies - with the communication of expert opinions. Furthermore, as suggested by Evett [18], harmonisation regarding verbal qualifiers aims at reducing problems of effective communication with mandating parties. It promotes exchange between forensic scientists of different disciplines as shown in the next section, where we present elements of the discussions that took place between scientists of our institution to address some of the challenges encountered when implementing a verbal scale.

2.3. Which type of scale should be used?

As underlined by the third principle of interpretation, the conclusions of our evaluative reports must refer to the probability of the observations given the propositions and the case information, and not to the probability of the propositions. We do not consider verbal equivalents based on probabilities of the propositions themselves to be appropriate, because assessment of the probability of the propositions falls within the competence of the Court.

Verbal equivalents based on probabilities of the propositions abound in literature. Examples can be found in [19,20] and in the "Standard terminology for expressing conclusions of forensic document examiners" [21], cited in the 2009 report of the National Research Council (p.166) [22]. Critiques of such proposals based on what is named posterior probabilities are given in [11,14,18,23-28]. As long as prior odds for the propositions are not specified, such scales, which are currently in use in some laboratories, amounts to a common fallacy known as transposing the conditional [29,30]. This is not the case for the scales proposed for example by Köller *et al.* [20] and the ENFSI Expert Working Group (EWG) Marks Conclusion Scale Committee [31], where equal prior odds are explicitly specified and adopted. However, again, we believe that prior odds should be specified by the Court, not by the experts [32]. Indeed, equal prior odds that are not based on the case at hand can lead to a result - that is posterior probabilities - that does not reflect the belief of the Court (or other recipients of expert information) [27,33]. It has been shown that the relevant figure for updating belief is the likelihood ratio (LR), not a posterior probability. However to help communicating the impact of the LR in a given case, we will suggest later to present various scenarios (see appendix C) among them the situation where prior probabilities are equal. Our objection to the approach by Köller *et al.* [20] and the ENFSI EWG Marks Conclusion Scale Committee [31] is not on the logical principles underpinning the calculation but on the choice to adopt, on behalf of the fact-finder, only one possibility for the prior probabilities.

It is for these reasons that the verbal expressions we consider relevant are focused on forensic observations and are related to likelihood ratio values, such as those in use in the former Forensic Science Service [4,18,34-36) or the Swedish National Forensic Centre [37], mentioned as an example also in the ENFSI guideline [2]. All these verbal scales are based on a convention and in essence there are no scales that are better than others [38]. What is important however is that all disciplines within a laboratory report the value of their observations using the same convention.

Theoretically, likelihood ratios range from 0 to infinity, but depending on the discipline or even the type of analysis, one may report different order of magnitudes: for example when performing DNA autosomal analysis, one can report likelihood ratios as large as a billion while with non autosomal (YSTR or mtDNA), likelihood ratios are generally much smaller. For micro-traces (such as glass, fibres or paint), LRs are also rarely larger than 10'000, but this may not be the case for other disciplines. As there are not an infinite number of verbal qualifiers, one has to choose how to slice the full range of reported LRs to construct a verbal scale. The ranges and the number of levels chosen should satisfy the needs of all reporting scientists in a given institution whatever the discipline. We therefore decided to propose a verbal scale similar to that of Evett *et al.* [4] using the following orders of magnitude: 1, 10, 100, 1000 and 10'000 and more. At the upper scale, because it is difficult to express very large LRs with words, the same word will describe different orders of magnitude (e.g., 10'000 and one billion). The difficulty of expressing large numbers with words is quite general, for example very rich people are described as billionaires or millionaires depending on their fortune. One does not really need to add that a billionaire is a very rich person, this is conveyed by the number itself, and this is why this number (or order of magnitude) should appear in the statement. The conventional scale we propose is illustrated in Appendix A. It is based on likelihood ratios and suitable for all disciplines.

2.4. Is the value of evidence properly conveyed if we use words only?

Implementing a unified verbal scale in an institution such as ours raises the question of whether the one we propose has the potential to improve communication between scientists and lay people and whether the recipient of information understands the results as intended. In reviewing the literature we found that even if it is widely considered that verbal terms based on the likelihood ratio are "the most appropriate basis for communication of an evaluative expert opinion to the court" (p.1) [39], there is empirical evidence for discrepancies between intentions in experts' conclusions and understanding drawn from such conclusions by recipients of expert information [40].

In the previous section we highlighted the importance of proposing a verbal scale on the evidence instead of one on the propositions. However, in the experiment reported by Sjerps and Biesheuvel [41], different verbal scales were presented to lawyers to study their preferences and assess their perception of probative value. Results showed that none of the lawyers detected the illogicality of one of the proposed scales that was based on propositions themselves. It is, thus, important to alert the recipient of information that to assess posterior odds (or the posterior probabilities of propositions) one needs to be given their prior odds (or prior probabilities of propositions).

Various research concentrated on how the strength of different verbal equivalents is perceived among people and on how this corresponds with the expert' results. Nordgaard *et al.* [42] noted that people tend to have their own interpretation of the verbal terms used by experts to communicate probative strength. It appears that the meaning of a single word depends on the context [43,44] and in particular on the evidence type, as some forensic types are expected to provide stronger evidence than others [45]. In our internal discussions on this topic, we have noted that it is sometimes disturbing to see that people think that likelihood ratios are relative in aspects that are not relative. For example, a LR of a 1000 in

glass expresses the same value (of probative strength) as a LR of 1000 in DNA. Therefore, one **cannot** say that a LR of 1000 is strong in glass and of limited value for autosomal DNA. And, on the other hand, people also sometimes think of LRs being absolute in aspects that are relative. For example, our LRs depend strongly on the propositions and on the case information.

The studies of Brun and Teigen [46] and Shaw and Dear [47] highlighted large differences between people in the attribution of a numerical equivalent to the same verbal expression. This result was confirmed by Martire *et al.* [44] who found a poor correspondence between verbal and numerical equivalents concerning the scale proposed by the Association of Forensic Science Providers (AFSP). This was also observed by McQuiston-Surrett and Saks [48] who examined undergraduate psychology students' ratings of the strength of the terms proposed by the American Board of Forensic Odontology (ABFO). They found that the responses from students were completely opposite to what the ABFO intended.

More recently, Martire and Watkins [40] intended to re-examine the results of the pilot study of Mullen *et al.* [45]. As in previous studies, the authors found that the verbal scale does not achieve its primary aim of facilitating communication between experts and lay people, because the perceived value of the forensic findings by lay people did not correspond to the intended value assigned to the evidence by experts. The authors analysed the correspondence between lay people interpretation of verbal qualifiers and ranges of LRs intended by the expert. According to their results, none of the ranges of LRs chosen by the experts corresponded with lay people perception of the strength of the evidence when presented with a verbal equivalent, except for "limited" or "weak support". However, even in those cases, Martire *et al.* [44] reported that the intentions of the expert could be misunderstood. The authors called this the "weak evidence effect", that is the risk of wrongly concluding that the evidence rather strongly supports the alternative proposition, in cases where the observations provide weak (or limited) support in favour of a given proposition over the alternative.

In order to improve the communication between experts and laypeople the next sections will be focusing on some solutions that could possibly overcome the above-mentioned difficulties.

2.5. Is there merit in providing the full range of the verbal scale (with or without levels) in the written statement?

It has been suggested in the literature that to help understand verbal expressions of probative strength, there would be merit in mentioning, in the scientist's written statement, all verbal expressions agreed and used by the expert's laboratory. This would enable the reader to see whether the case at hand should be considered 'a strong case', or on the contrary, a case where the evidence is of limited value. This view is recommended, for example, by the AFSP statement [1], which requires that the full range of verbal expressions be provided in the expert's report. A position that is also supported by Jackson *et al.* [49]. If, as a first step to help implementation of the approach, one decides not to express the likelihood ratio in terms of numbers, then we see advantages in providing the whole scale in the written statement. However, we argue that mentioning the full range of verbal expressions comes with some costs and appears to be of limited value for harmonising the way in which the strength of evidence is understood. Indeed, in the study of Sjerps and Biesheuvel [41], the value of the observations was perceived differently among different persons, *even when* the various verbal terms were fully disclosed to participants. A further issue is that stating the full range of verbal qualifiers may suggest that a likelihood ratio falling in the top level of the range is more 'useful' for the fact-finder than a likelihood ratio from a lower level. This conveys the misleading impression that the LR obtained in the given case is relative to other cases one

could possibly have. For these reasons, we do not recommend to provide the full range of verbal qualifiers in the written statement.

A topic related to the above is the use of the log-likelihood ratio (log(LR)) [23,50,51]. As early as 1950, Good [50] referred to the log(LR) as the *weight of evidence*, because it has several desirable properties that a measure of information ought to have. Most prominently known is the additive property, that is the joint value of several probabilistically independent items of evidence is given by the sum of the weights (i.e., log-likelihood ratios) attached to each item of evidence. Furthermore, one can sum the weight of evidence with the logarithm of the prior odds to obtain the logarithm of the posterior odds. This has the advantage of keeping the intuitive image of weighing evidence in the scales of justice: when the weight is 0 the evidence provides no support and the scales of justice remain unchanged; when the weight is positive it provides support for the first proposition and when it is negative it provides support for the alternative. Finally, using the log(LR) helps understand the magnitude of large likelihood ratio values. Different commonly used scales refer to logarithms for such reasons: the Richter scale for earthquakes, decibels for sound, pH for acidity.

In our scale, however, we decided not to indicate the log(LR) because the use of numbers such as +1, +2 etc. to indicate the log(LR) associated to the verbal qualifiers might lead one to think that a +1 is not useful, and that one needs greater tags, such as +5, in order to have powerful evidence. This, again, provides the inappropriate impression that the likelihood ratio in the given case is relative to other possible cases. However, recipients of expert information are (or should be) exclusively concerned by the likelihood ratio obtained in their case at hand. They will consider their case as a whole and ideally combine their prior beliefs about the case with the value provided by scientific observations. For example, if their prior odds for the main proposition in the case are high, then even a small likelihood ratio might be useful in the sense of being sufficient to discriminate between the propositions of interest and thus help with the case issue. On other occasions, even a large likelihood ratio may not be helpful enough because prior odds are very low.

We concede that the wording used for verbal qualifiers may also lead to the same perception. A value of observations qualified as providing ‘weak (or limited) support’ for a given proposition over the other may suggest that the evidence is quite useless. However, it may be very helpful for the Court, depending on their prior odds. In our next section, we discuss ways to tackle this challenge.

2.6. What options can we explore to improve communication?

In the previous paragraphs, we have seen that one disadvantage of verbal qualifiers is that they may lead to the spurious view according to which the value of the observations made in a given case is relative to other cases. Verbal qualifiers are also prone to misunderstandings and cannot be coherently combined with other evidence. For these reasons, we should not report the associated verbal qualifier alone: one can either give one's LR and the verbal qualifier or just give the LR alone. Reporting only a numerical value would seem to be the best option as numbers can be logically combined [12,23]. However, we have also seen that people generally tend to be poorly acquainted with matters concerning the measurement of uncertainty using probability. Before key players are fully accustomed to likelihood ratios as a measure of the value of evidence, communication needs to be improved. Hereafter we suggest a few avenues. These are of course non prescriptive and the best option may depend on personal preference of either the scientist or the recipient of information or even on the case.

Because recipients of information are *in fine* interested in posterior probabilities, we suggest showing the impact of the value of evidence on the probability of the propositions. Therefore,

we suggest to illustrate this impact with a table where one takes several prior probabilities (or odds), and then provide the corresponding values of posterior probabilities (or odds) given the LR obtained in the case.

For illustration purposes, we have taken a case with a likelihood ratio in the order of ten. In such a case, we would qualify the value of the observations as weak or limited. Indeed, as mentioned earlier, this type of value has been shown to be inappropriately conveyed by the use of verbal equivalents alone. To avoid the problem that laypersons believe that limited support for one proposition means, for example, strong support for the alternative proposition (also referred to the weak evidence effect [44]), we prefer to formulate the conclusion in two steps. First, the forensic scientists could state that the observations support a given proposition over the other. Then, in a second sentence, they would qualify this support as limited or weak. Moreover, we emphasise that it is crucial to always mention both propositions in our conclusions, and not only one as the value of the results depends intimately on the propositions at hand. This procedure has the advantage of underlining that the value of the results obtained is not absolute, but relative to the propositions of interest. An example of such a conclusion is provided in Appendix B. In Appendix C, Tables 2 and 3 show the impact of the likelihood ratio obtained on posterior odds (and posterior probabilities of defence proposition), taking three examples of prior odds (and prior probabilities of defence proposition).

As we have shown before, scientists must report their conclusions based on the probability of the observations given the propositions. However, the recipient of information may misunderstand this conclusion as an opinion on the propositions themselves. This is the above-mentioned well-known error called transposing the conditional. In order to avoid such a fallacy, one can indicate in the statement not only what the results mean but also what they do not mean. An example is given in appendix D. In particular, if the statement must be translated, such a word of caution may be helpful in order to ensure that the translator does not transpose the conditional either.

3. Conclusion

The findings of the studies of Sjerps and Biesheuvel [41], Mullen *et al.* [45] and Martire and Watkins [40] show that it is not sufficient to develop only appropriate terminology for verbal terms. Improvement of education in this area is also essential. The focus should be on explaining in detail the experts' logic of reasoning, and the way in which it contributes to reasoning at trial in general. It appears less promising to concentrate on the likelihood ratio alone, since it is not well understood in isolation, in whatever way it is communicated, that is a number, a verbal equivalent or even visually [44].

This avenue, as well as other initiatives such teaching, conferences, articles, continuing education [52], research and discussions between scientists and the recipients of information need to be pursued in order to improve communication and mutual understanding. A verbal equivalent alone cannot achieve the difficult task of helping the Court fully grasp the concept of probability and likelihood ratios.

Verbal qualifiers for probative strength, in whatever way they are agreed and practiced within the forensic community, or within a forensic institution, merely represent a consensus. However, they are widely found to be unsatisfactory, both for the general audience (since no standardisation is possible regarding the individual understanding of verbal equivalents) and scientists. Indeed, there does not seem to be a single right word that would express a given aspect of expert opinion. For example, Martire *et al.* [44] have pointed out that the verbal conventions proposed by the AFSP are far from ideal.

A fundamental question that follows is whether verbal equivalents could be abandoned completely if (i) experts were able to coherently articulate numbers, including situations in which they are based on experience and scarce 'hard' data, and (ii) recipients of expert information would be able to properly handle such numbers. An immediate answer would be 'yes', however numerical expressions of probative value are usually not appropriately understood - contrarily to other numerical measures that most of us understand without the need of any further explanation (e.g., expressions for length, temperature, etc.). A more widespread understanding of the likelihood ratio as *the* measure of probative value thus appears desirable, but as long as this is not the case, we consider it necessary to follow the current ENFSI guideline and rely on both the numerical likelihood ratio and the option of verbal qualifiers, where deemed useful, for evaluative reporting in forensic science.

Acknowledgments

We gratefully thank the participants who took an active part in the fruitful discussions that provided the substance of this paper. They constitute a rich panel of practitioners coming from various forensic disciplines – such as questioned documents, fingerprints, DNA, microtraces, shoemarks – as well as specialists of forensic interpretation. Most participants were internal forensic scientists of the School of Criminal Justice of the University of Lausanne. Some participants were members from the following Swiss institutions: the *Service Forensique* of Neuchâtel, especially Dr Simon Baechler; the Forensic Science Institute of Zurich, represented by Mr Erich Kupferschmid, as well as the AFIS DNA Services of the Federal Office of Police, directed by Dr Axel Glaeser. We also express our gratitude to Charles Berger, who made us aware of the fact that people tend to think that the likelihood ratio is relative when it is not, and that it is not relative when it is! Finally we acknowledge the reviewers for their very useful comments.
Alex Biedermann gratefully acknowledges the support of the Swiss National Science Foundation through grant No. BSSGI0_155809.

APPENDICES

A. Proposed verbal scale

A likelihood ratio in excess of one represents support for the prosecution proposition. A likelihood ratio less than one represents support for the defence proposition. It cannot support both propositions, but it can support neither, when the likelihood ratio is equal to one. Indeed, if the results are just as probable given the truth of either proposition, they do not help to advance the issue.

However, as LRs below one are difficult to understand, we found it beneficial to inverse the propositions if the support is in favour of the defence. Therefore, instead of reporting that – given the information available – the results are 0.1 times more probable given the prosecution proposition than given the defence proposition, we will write that – given the information available – the results are 10 times more probable given the defence proposition than given the prosecution proposition.

VERBAL COMMUNICATION	LR
The results support the proposition that... rather than the proposition that… This support is qualified as *extremely strong*.	**> 10'000**
The results support the proposition that... rather than the proposition that... This support is qualified as *very strong*.	**>1000 – 10'000**
The results support the proposition that... rather than the proposition that... This support is qualified as *strong*.	**>100 – 1000**
The results support the proposition that... rather than the proposition that... This support is qualified as *moderate*.	**>10 – 100**
The results support the proposition that... rather than the proposition that... This support is qualified as *weak or limited*.	**>1 – 10**
The results support neither propositions. This support is qualified as *null*.	**1**

Table 1: Proposed verbal scale for reporting the value of the scientific observations (translated from French).

B. Example of how to formulate a conclusion when our LRs are in the order of 10 (i.e., defined as weak evidence).

It has been shown in the literature [44] and in practice that verbal qualifiers chosen to convey small LRs are not well understood. It is hoped that by reporting in two phases, the value of the observations will be communicated more efficiently. We have taken a real case where the propositions were:

- Mr Jones signed the contested document (prosecution proposition);
- An unknown person signed the contested document (defence proposition).

As our LR was in the order of 0.1, we would reverse the propositions and report as follows: *The results provide support for the proposition that an unknown person – rather than Mr Jones – signed the contested document. This support is qualified as weak or limited, as the results are in the order of 10 times more probable given that the proposition that an unknown person signed the contested document is true, rather than given that the alternative is true (i.e., Mr Jones signed the contested document).*

C. Example of how one can show the impact of the value of the observations in the case at hand

The relevant figure for updating belief is the likelihood ratio, not a posterior probability. However, how the Court is to use the information provided by the forensic examinations and combine it with the other elements of the case is a challenge. There are certainly several ways to deal with this aspect, but one possible solution could be to present, in the form of an appendix to the statement, the impact of the LR obtained in the case at hand on posterior odds, based on different prior odds. Here is an example taking again the previous signature case (with a likelihood ratio of 10 in favour of defence proposition).

The impact of the value of the observations made on the signatures on the probability of a proposition is illustrated in Table 3. In order to obtain the posterior probability (considering the information provided in the case *and* the forensic observations) that an unknown person has signed the document rather than Mr Jones, one has to combine the likelihood ratio (i.e., the value of the observations provided by the document examiner) with prior odds (considering all other relevant element in the case but the forensic observations) that an unknown person has signed the document rather than Mr Jones. These prior odds – or respectively one's prior probability – represent the opinion that the Court may have on the probability that an unknown person (and respectively the probability that Mr Jones) signed the document based on the other elements of the case, before being exposed to the results of the forensic document examination. The posterior probability represents the opinion that the Court would have on the proposition that an unknown person signed the document based on the other elements of the case *and* on the findings obtained by the forensic document examiner.

	Prior odds for defence proposition compared to prosecution	**LR**	**Posterior odds for defence proposition compared to prosecution**
Example 1	1 to 9	10	10 to 9
Example 2	1 to 1	10	10 to 1
Example 3	9 to 1	10	90 to 1

Table 2: Examples of posterior odds obtained by multiplication of prior odds by our likelihood ratio. Using odds, one can easily see that – with a LR of 10 – posterior odds are ten times larger than prior odds.

Persons who would rather use probabilities than odds can refer to Table 3.

	Prior probability for defence proposition	**LR**	**Posterior probability for defence proposition**
Example 1	0.1	10	0.5
Example 2	0.5	10	0.9

Example 3	0.9	10	0.99

Table 3: Examples of posterior probabilities considering different prior odds and a likelihood ratio of 10. To obtain our posterior probabilities, we used the odds form of Bayes formula and calculated the corresponding probabilities [3].

One can see that the posterior probability depends not only on the observations made in the context of the forensic examination but also on the value that the Court's prior probability may take (which depends on the other elements of the case).

Three cases are illustrated, but on demand it is possible to consider more:

- The first case illustrates the situation where the balance of prior probabilities favours the proposition that Mr Jones signed the document (with a probability of 0.9) rather than some unknown person (who would have signed with a probability of 0.1). On one hand, one can see that after the information provided by the signature examination, the probability that some unknown person signed goes from 0.1 to 0.5. In that case after considering the findings, it is equally probable that Mr Jones has signed rather than an unknown person.
- The second case illustrates the situation where the balance of prior probabilities favours neither proposition (both are *a priori* equally probable). One can see that after knowing the results of the forensic examination, the posterior probability that some unknown person signed goes from 0.5 to 0.9. On the other hand, the posterior probability that Mr Jones signed the document goes from 0.5 to 0.1.
- The third case illustrates the situation where the balance of prior probabilities favours the proposition that an unknown person signed the document (with a probability of 0.9) rather than Mr Jones (who would have signed with a probability of 0.1). After considering the information given by the examination of the signatures, the posterior probability that some unknown person signed the contested document goes from 0.9 to 0.99. On the other hand, the posterior probability that Mr Jones signed the document goes from 0.1 to 0.01.

D. Note of caution regarding transposing the conditional

As recipients of information are interested in the probability of the propositions, conclusions such as "The results strongly support the prosecution proposition rather than the defence proposition" are often read (or translated) as "Prosecution proposition is highly probable". To avoid such a misunderstanding, one can put a cautionary note in the statement that can take the following form:
Note of caution: Our results do not mean that it is probably an unknown person who signed the document. Indeed, the probability that it is an unknown person (rather than Mr Jones) who signed the contested document depends not only on the observations made on the signatures, but also on other elements (enquiry, testimony, other information). The evaluation of these other elements are of the domain of the Court, and scientists should not give their opinion on the truth (or probability) of the propositions, but they should help the trier of facts by giving their probability of the observations given each proposition.

References

[1] AFSP (Association of Forensic Science Providers) Standards for the formulation of evaluative forensic science expert opinion, *Science and Justice* 49 (2009) 161-164.

[2] Willis S., Mc Kenna L., Mc Dermott S., O' Donnell G., Barrett A., Rasmusson B., Höglund T., Nordgaard A., Berger C., Sjerps M., Molina J. J. L., Zadora G., Aitken C., Lovelock T., Lunt L., Champod C., Biedermann A., Hicks T., Taroni F. ENFSI Guideline for evaluative reporting in forensic science, 2015.
www.enfsi.eu/news/enfsi-guideline-evaluative-reporting-forensic-science

[3] Robertson B., Vignaux G. A. Interpreting evidence: Evaluating forensic science in the courtroom, John Wiley and Sons Ltd, Chichester, United Kingdom, 1995.

[4] Evett I. W., Jackson G., Lambert J. A., McCrossan S. The impact of the principles of evidence interpretation on the structure and content of statements, *Science and Justice* 40 (2000) 233-239.

[5] Aitken C. G. G., Taroni F. Statistics and the evaluation of evidence for forensic scientists, 2nd Edition, John Wiley and Sons Ltd, Chichester, United Kingdom, 2004.

[6] Kind S. S., Wigmore R., Whitehead P. H., Loxley D. S. Terminology in forensic science, *Journal of the Forensic Science Society* 19 (1979) 189-191.

[7] Brown, G. A., Cropp, P. L. Standardised nomenclature in forensic science, *Journal of the Forensic Science Society* 27 (1987) 393-399.

[8] Leung, S. C., Cheung, Y. L. On opinion, *Forensic Science International* 42 (1988), 1-13.

[9] Evett I. W. Bayesian inference and forensic science: Problems and perspectives, *Journal of the Royal Statistical Society. Series D (The Statistician)* 36 (1987) 99-105.

[10] Jackson G. The scientist and the scales of Justice, *Science and Justice* 40 (2000) 81-85.

[11] Champod C., Evett I. W., Jackson G., Birkett J. Comments on the scale of conclusions proposed by the *ad hoc* committee of the ENFSI marks working group, *Information Bulletin for Shoeprint/Toolmark Examiners* 6 (2000) 11-18.

[12] Lindley D. The philosophy of statistics, *Journal of the Royal Statistical Society. Series D (The Statistician)* 49 (2000) 293-337.

[13] Robertson B., Vignaux G. A., Berger C. E. H. Extending the confusion about Bayes, *The Modern Law Review* 74 (2011) 444-455.

[14] Berger C. E. H., Buckleton J., Champod C., Evett I. W., Jackson G. Evidence evaluation: A response to the court of appeal judgment in R *v* T**,** *Science and Justice* 51 (2011) 43-49.

[15] Balding D. J., Steele C. D., Weight-of-evidence for forensic DNA profiles, 2nd Edition, John Wiley and Sons Ltd, Chichester, United Kingdom, 2015.

[16] Stoney D. A. Evaluation of associative evidence: Choosing the relevant question, *Journal of the Forensic Science Society* 24 (1984) 473-482.

[17] Faigman D., Jamieson A., Noziglia C., Robertson J., Wheate R. Response to Aitken et al. on R v T, *Science and Justice* 51 (2011) 213-214.

[18] Evett I. W. Towards a uniform framework for reporting opinions in forensic science casework, *Science and Justice* 38 (1998) 198-202.

[19] McAlexander T. V., Beck J., Dick R. M. The standardization of handwriting opinion terminology, *Journal of Forensic Sciences* 36 (1991) 311-319.

[20] Köller N., Niessen K., Riess M., Sadorf E. Probability conclusions in expert opinions on handwriting, substantiation and standardization of probability statements in expert opinions, Luchterhand, München, 2004.

[21] ASTM Standard terminology for expressing conclusions of forensic document examiners, Designation E 1658-08.

[22] NRC Strengthening Forensic Science in the United States: A Path Forward, The National Academies Press, Washington D.C., 2009.

[23] Champod C., Evett I. W. Commentary on A. P. A. Broeders (1999) 'Some observations on the use of probability scales in forensic identification', Forensic Linguistics 6(2): 228–241, *International Journal of Speech Language and the Law* 7 (2000) 238-243.

[24] Evett, I. W. Verbal convention for handwriting opinions, *Journal of Forensic Sciences* 45 (2000) 508-509.

[25] Taroni F., Aitken C. G. G. Fibres evidence, probabilistic evaluation and collaborative test, *Forensic Science International* 114 (2000) 45-47.

[26] Taroni F., Buckleton J. Likelihood ratio as a relevant and logical approach to assess the value of shoeprint evidence, *Information Bulletin for Shoeprint/Toolmark Examiners* 8 (2002) 15-25.

[27] Taroni F., Biedermann A. Inadequacies of posterior probabilities for the assessment of scientific evidence, *Law, Probability and Risk* 4 (2005) 89-114.

[28] Biedermann A., Taroni F., Aitken C. G. G. Letter to the editor – re: Conclusion scale for shoeprint and toolmarks examinations", *Journal of Forensic Identification* 56 (2006) 685-689.

[29] Thompson W. C., Schumann E. L. Interpretation of statistical evidence in criminal trials: The prosecutor's fallacy and the defence attorney's fallacy, *Law and Human Behavior* 11 (1987) 167-187.

[30] Evett I. W. Avoiding the transposed conditional, *Science and Justice* 34 (1995) 127-131.

[31] ENFSI Expert Working Group Marks Conclusion Scale Committee, Conclusion scale for shoeprint and toolmarks examination, *Journal of Forensic Identification* 56 (2006) 255-280.

[32] Thompson W. C., Vuille J., Biedermann A., Taroni F. The role of prior probability in forensic assessments, *Frontiers in Genetics* 4 (2013) Article 220. http://journal.frontiersin.org/article/10.3389/fgene.2013.00220/full

[33] Biedermann A., Taroni F., Garbolino P., Equal prior probabilities: Can one do any better? *Forensic Science International* 172 (2007) 85-93.

[34] Cook R., Evett I. W., Jackson G., Jones P. J., Lambert J. A. A model for case assessment and interpretation, *Science and Justice* 38 (1998) 151-156.

[35] Stockton A., Day S. Bayes, Handwriting and Science, presentation given at the *Meeting of the American Society of Questioned Document Examiners (ASQDE)*, 2001, Des Moines, Iowa.

[36] Jackson G., Jones P. J. Case assessment and interpretation, Wiley Encyclopedia of Forensic Science, Allan Jamieson, André Moenssens Eds., 2009.

[37] Nordgaard A., Ansell R., Drotz W., Jaeger L. Scale of conclusions for the value of evidence, *Law, Probability and Risk* 11 (2012) 1-24.

[38] Champod C., Baldwin D., Taroni F., Buckleton J. Firearm and tool marks identification: The Bayesian approach, *AFTE Journal* 35 (2003) 307-316.

[39] Evett I. W. and other signatories Expressing evaluative opinions: A position statement, *Science and Justice* 51 (2011) 1-2.

[40] Martire K. A., Watkins I. Perception problems of the verbal scale: A re-analysis and application of a membership function approach, *Science and Justice* 55 (2015) 264-273.

[41] Sjerps M., Biesheuvel D. B. The interpretation of conventional and Bayesian verbal scales for expressing expert opinion: a small experiment among jurists, *Forensic Linguistics* 6 (1999) 214-226.

[42] Nordgaard A., Widstedt I., Drotz W., Elmqwist J., Höglund T., Jaeger L., Torbjörnnsson M., Palmborg J., Sullivan S, Wigilius I. Uppfattning av vërdeord i sakkunnugztlatanden-En studie genomförd bland olika aktörer i rättsprocessen i Sverige, Swedish National Laboratory of Forensic Science, SKL Rapport, 2010.

[43] Mosteller F., Youtz C. Quantifying probabilistic expressions, *Statistical Science* 5 (1990) 2-12.

[44] Martire K. A., Kemp R. I., Watkins I., Sayle M. A., Newell B. R. The expression and interpretation of uncertain forensic science evidence: Verbal equivalence, evidence strength, and the weak evidence effect, *Law and Human Behavior* 37 (2013) 197-207.

[45] Mullen C., Spence D., Moxey L., Jamieson A. Perception problems of the verbal scale, *Science and Justice* 54 (2014) 154-158.

[46] Brun W., Teigen K.H. Verbal probabilities: ambiguous, context-dependent, or both? *Organizational Behaviour and Human Decision Processes* 41 (1988) 390-404.

[47] Shaw N., Dear P. How do parents of babies interpret qualitative expressions of probabilities? *Archive of Disease in Childhood* 65 (1990) 520-523.

[48] McQuiston-Surrett D., Saks M.J. Communicating opinion evidence in the forensic identification sciences: Accuracy and impact, *Hastings Law Journal* 59 (2008) 1159-1189.

[49] Jackson G., Evett I. W., Champod C., Buckleton J. Letter to the Editor – re: Perception problems of the verbal scale, *Science and Justice* 54 (2014) 180.

[50] Good I. J. Weight of evidence and the Bayesian likelihood ratio, in The use of statistics in forensic science (Eds: Aitken C. G. G. and Stoney D. A.). Ellis Horwood, Chichester, UK, 1991, pp. 85-106.

[51] Aitken C. G. G., Taroni F. A verbal scale for the interpretation of evidence, *Science and Justice* 38 (1998) 279-281.

[52] Biedermann A., Hicks T., Voisard R., Taroni F., Champod C., Aitken C. G. G., Evett I. W. E-learning initiatives in forensic interpretation: Report on experiences from current projects and outlooks, *Forensic Science International* 230 (2013) 2-7.