Science and Justice 55 (2015) 264–273

Contents lists available at ScienceDirect

Science and Justice

journal homepage: www.elsevier.com/locate/scijus

Perception problems of the verbal scale: A reanalysis and application of a membership function approach

Kristy A. Martire *[,1], Ian Watkins

School of Psychology, The University of New South Wales, Australia

ARTICLE INFO

Article history:
Received 30 July 2014
Received in revised form 18 December 2014
Accepted 5 January 2015

Keywords:
Verbal conclusion scales
Likelihood ratios
Communicating uncertainty
Membership functions
Expert evaluative opinions

ABSTRACT

Verbal conclusion scales provide a standardized vocabulary that forensic scientists can adopt to describe the amount of support offered by a set of observations with regard to two competing hypotheses. The extent to which these verbal scales can efficiently and accurately communicate strength of support to lay evaluators is, however, an empirical matter of considerable importance. The aim of this paper was to reexamine the results of a recent study measuring lay interpretations of expert verbal phrases (Mullen et al., 2014) and to further improve upon those estimates through the utilization of a membership function approach. Across both the reexamination (n = 400) and the new experiment (n = 134) 534 participants provided translations of expert verbal conclusion scales used by forensic scientists. Overall, there is compelling evidence that the correspondence between expert intentions and lay interpretations is low, while the potential for miscommunication is high. Consequently, further attention is required to facilitate the development of valid and reliable verbal conclusion scales which clearly communicate expert evaluative opinions.

While it is clear that many within the forensic science community have been grappling with issues regarding the standardization of expert evaluative opinions over a number of years [2–6], the 2009 National Academies of Science (NAS) report has brought these considerations into sharp relief and has provided authority for reform [7]. Indeed, as a consequence of the NAS having demarked communication as a priority issue in the forensic sciences the structure and content of expert conclusions has been rapidly evolving and extensively discussed in recent years [8–11]. Specifically, authoritative individuals and organizations have garnered momentum for their move away from conclusions regarding individualization that are expressed through discipline-specific verbal labels [12], and toward the use of likelihood ratios and uniform verbal numerical translations as best practice for the communication of evidential weight [4,13–15].

The UK Association of Forensic Science Providers (AFSP) is one group who has published verbal–numerical translations for the purpose of standardizing and simplifying the communication of evidence strength in likelihood ratio format. In their paper *Standards for the formulation of evaluative forensic science expert opinions* the AFSP provided an example of a scale currently employed by "many practitioners" [16; see Table 1] and advised that the scale will be adopted by all AFSP organizations. Questions have arisen, however, regarding the extent to which the suggested verbal expressions of uncertainty serve the purpose of effectively and accurately [13] communicating opinions regarding evidential weight to the lay consumer (i.e., lawyers, judges and jurors).

Any questions regarding correspondence between expert intentions and lay interpretations of evaluative opinions can be considered in the context of psychological research examining how people understand phrases describing uncertainties. Broadly speaking the evidence suggests that decision-makers struggle to extract the intended information from verbal labels describing uncertain outcomes. For example, studies where participants have been asked to provide a numerical translation for a probabilistic statement like *almost certain*, *likely*, or *doubtful*, have revealed that the meaning of phrases are quite vague and extend over a range of probabilities [17] rather than communicating a precise interpretation.

This is consistent with a body of literature showing considerable intra- and inter-individual variation in the understanding and use of probability phrases generally [17], and by expert forensic scientists [18,19]. This means that vagueness in interpretations can be seen within individuals as well as across individuals, leaving probability phrases as diffuse rather than precise communication tools. However, although the vagueness of probability terms and the potential for variable interpretations have been widely observed across a range of decision-making contexts, from climate change [20,21] to jury decision making [22] and beyond [23,24], it is also the case that this literature has a somewhat limited application to our understanding of the verbal expressions proposed by the AFSP. This is because the AFSP verbal expressions aim to communicate a degree of *support*, rather than a degree of *uncertainty*.

* Corresponding author at: School of Psychology, The University of New South Wales, Sydney, NSW 2052, Australia.
E-mail address: k.martire@unsw.edu.au (K.A. Martire).
[1] ORCID: http://orcid.org/0000-0002-5324-0732.

http://dx.doi.org/10.1016/j.scijus.2015.01.002
1355-0306/

Table 1
Examples of expert verbal conclusion scales.

Value of likelihood ratio	Verbal equivalent (support for proposition)
>1–10	Weak[a]/weak or limited[b]/limited[c]
10–100	Moderate[a–c]
100–1000	Moderately strong[a–c]
1000–10,000	Strong[a–c]
10,000–1,000,000	Very strong[a–c]
>1,000,000	Extremely strong[a–c]

[a] Association of Forensic Science Providers [16].
[b] Martire et al. [28]; Martire et al. [29].
[c] Mullen et al. [1].

Opinions verbally or numerically describing the evidential weight offered by a set of observations in a likelihood ratio format, like those advocated for by the AFSP and others [14,25–27], describe the ratio of the probability of a set of observations given one (prosecution) hypothesis relative to the probability of the observations given a competing (defense) hypothesis [13]. This *ratio of probabilities* constitutes an estimate of the *magnitude of support* that the observations can offer for one hypothesis as compared to the other. Strictly speaking, it is not a communication of uncertainty in and of itself, nor is it a communication regarding the uncertainty about the observations given either of the propositions (e.g., it is *highly likely* that we would make this observation if the prosecution's hypothesis were true, but it is *highly unlikely* that we would make this observation if the defense's hypothesis were true). Rather, lay decision-makers presented with a verbally constituted likelihood ratio are being provided information about the amount of support being offered for one proposition compared to the other (e.g., the findings provide *strong* support for the prosecution hypothesis rather than the defense hypothesis). Unfortunately, the broader psychological literature is relatively silent on how lay decision-makers will interpret verbal labels communicating the amount of support provided one hypothesis as compared to another based on the ratio of assigned probabilities, as in the case of the likelihood ratio.

To these authors knowledge just three papers to-date have directly examined verbal scales communicating the strength of support provided by the ratio of two probabilities. Before discussing each of these in detail it is worth noting that the high degree of similarity between the scales tested appears to have been the source of some confusion regarding their attribution to source in two instances. Martire et al. [28,29], attributed the scale that they examined (see Table 1) to the Association of Forensic Science Providers [16]. However it differs from that of the AFSP in one small respect — the verbal equivalent used for the lowest level of support (LR range >1–10) "weak or limited support" comes from the scale used in *R v T* [30] rather than the AFSP who specified "weak support". The two scales are otherwise identical in every respect. A third scale was examined by Mullen et al. [1; see Table 1]. This scale also differed from the AFSP scale with respect to the verbal label used for the lowest level of support — in this case the phrase "limited support" was used rather than "weak support". Despite their differences, together these studies highlight some of the variability that forensic scientists are seeking to minimize in relation to the communication of evaluative expert opinions, while also providing some insight into lay interpretations of their meaning.

In their first paper Martire et al. [28] reported two experiments examining the equivalence of the verbal labels and the numerical values specified in the conclusion scale shown in Table 1. In the first experiment, 494 online respondents were presented a set of case facts in a hypothetical larceny trial before being asked to estimate their prior-belief in the guilt of the defendant. Participants were then presented an evaluative opinion from an expert shoe mark analyst whose testimony was provided in the form of either a verbal or a numerical likelihood ratio with regard to a prosecution and a defense hypothesis of the crime. The strength of the support offered by the evidence was also varied such that participants were advised that the shoe mark comparison offered either: weak or limited support (LR = 4.5), moderately strong support (LR = 450), or very strong support (LR = 495,000) for the prosecution as compared to the defense hypotheses. After receiving this information, participants were asked to provide their posterior-belief in the guilt of the defendant. The difference between the posterior- and prior-beliefs constituted a measure of the weight attributed to the evidence by the lay evaluator.

The experiment revealed two main findings. First, numerical and verbal expressions resulted in statistically equivalent amounts of belief-change when the evidence strength was moderate or high. By contrast, when evidence strength was low only the verbal expression *weak or limited* caused participants to treat evidence which weakly supported the prosecution hypothesis as though it weakly supported the defense hypothesis. Such an effect, also known as a *weak evidence effect* suggested that participants could have been mistaking the direction of the support offered by the evidence.

The second important result was with regard to the amount of weight that participants attributed to the evidence when compared to the meaning intended by the expert. By comparing the strength of evidence interpreted by the participants (implicit likelihood ratios; ILRs), with the likelihood ratios provided by the expert it was possible to see a substantial undervaluing of the evidence with median ILRs orders of magnitude smaller than those intended in the moderate and high strength conditions. This was true for both verbal and numerical expressions although to a lesser extent in the numerical condition. Overall, these results led the authors to question whether the use of verbal equivalents should be endorsed for the presentation of low-strength evidence (*weak* support) and whether likelihood ratio formats were well understood by lay decision makers.

The results of the second experiment undertaken by Martire et al. [28] further showed the variability of the phrases when it was identified that simply changing the order of the propositions; that is, recasting the same experts' opinion regarding the evidence as being in support of the defense rather than the prosecution proposition, could change how the expression *weak or limited* was interpreted. Participants presented with the same weak evidence in Experiment 2 were no longer subject to a weak evidence effect when the weak evidence was exculpatory rather than inculpatory. Thus demonstrating that far from more closely approximating expert intentions, lay interpretations of the verbal support phrases can be influenced by contextual factors beyond the evidence itself.

In their second paper Martire et al. [29] trialed four different methods for communicating the support described by a likelihood ratio. In that experiment 404 online respondents assessed expert evaluative opinions regarding high or low strength evidence presented in one of four formats: numerical, verbal, table or visual scale. Similar to the approach taken in Martire et al. [28] those in the verbal and numerical conditions simply received an opinion of *weak or limited* support (LR 5.5) or *strong* support (LR 5500) in the low and high strength groups respectively. Those given the table format were presented with a table describing all possible verbal and numerical likelihood ratios contained within a verbal conclusion translation table like that published by the AFSP, with either the low or the high strength cells of the table highlighted according to condition. Finally, those in the visual scale condition were presented with an image of a line spanning from "in favour of hypothesis 2" on the left, to "in favour of hypothesis 1" on the right with a "neutral" midpoint. The value of the evidence was marked with regard to the two hypotheses by placing a red 'X' on the line in the appropriate location. Belief-change was again measured by eliciting prior- and posterior-beliefs regarding the guilt of the accused.

Unfortunately, the alternative methods did little to improve the correspondence between the expert intentions and lay interpretations. In keeping with past observations weak evidence effects were observed in the condition where verbal low strength evidence was presented. Moreover, a comparison of the implied likelihood ratios used by