# Response to “a study of the perception of verbal expressions of the strength of evidence”

Charles E.H. Berger[a,b], Reinoud D. Stoel [a]

[a)] *Netherlands Forensic Institute, PO Box 24044, 2490 AA, The Hague, The Netherlands.*

[b)] *Institute for Criminal Law and Criminology, Faculty of Law, Leiden University, PO Box 9520, 2300 RA Leiden, The Netherlands.*

We would like to respond to the recent paper “Understanding forensic expert evaluative evidence: A study of the perception of verbal expressions of the strength of evidence”, by Arscott et al. [1].

We agree that a verbal expression of the strength of evidence can be interpreted in varying ways. Not only by the people that read them, but also by those that express them. It is also possible that different verbal expressions are interpreted in the same way by different readers or reporting scientists. This is the reason that the Association of Forensic Science Providers (AFSP) and the European Network of Forensic Science Institutes (ENFSI) have published guidelines that call for forensic institutes to provide verbal scales and numerically define the verbal expressions therein [2, 3]. This, at least formally, solves the issue of the perception of intended strength of evidence.

If the intention is to convey the levels of support of verbal expressions, we wonder why participants were not provided with the intended levels of support. This is all the more surprising since the paper implies (p. 221) that a likelihood ratio (LR) is calculated, in which case there is no reason to use a verbal expression instead of the number calculated itself. While using verbal expressions is never necessary, they are an option when the LR is not the result of a calculation but the proper way to express a qualitative evaluative expert opinion. This is what happens in the majority of cases.

This study does not address the participants’ level of understanding of the *concept* of evidential strength. It implicitly assumes participants understood the concept of evidential strength (it models them as understanding the concept). This assumption is unfounded, because even the majority of forensic scientists themselves did not understand this concept until relatively recently. Many still don’t, and it is common to encounter examples of misinterpretation and misapplication of the concept of evidential strength. The fallacy of the transposed conditional, for example, is still being made in case reports, court transcripts, and the press.

This fallacy entails that one mistakenly perceives the expression of the strength of evidence as an expression of the probability that the proposition of the prosecution is true. Understanding this misconception is key to understanding some of the results of this study. We think that it may even explain the 'weak evidence effect' [4, 5], which is puzzling if you cling to the assumption that participants understood the concept of evidential strength.

If we model the perception of some of the participants as going through a transposition of the conditional, the 'weak evidence effect' is not so surprising anymore. If we have evidence that weakly supports the prosecution's proposition, and we commit this fallacy, it would seem that the prosecution's proposition has a low probability of being true. That would make the defence's proposition seen as very probably true (implicitly assuming exhaustivity). Transposing the conditional again, this would be seen as strong evidence for the defence's proposition. This problem is to be expected if one uses the language of evidential strength without fully understanding the concept. The 'weak evidence effect' may therefore not be the result of intrinsic problems of the verbal scale expression scale, but of the misunderstanding of the concept of evidential strength.

It is not clear to us why the participants would need to hear about a case if they were just to express how they perceived the expert's opinion on the strength of the evidence. The information about the case could only foster contextual bias and encourage them to replace or combine the expert's opinion with their own. The authors are well aware of the effect the context of the case may have on the outcome when they write "*A volume crime – a burglary – was selected as it was hypothesised that a more "serious" offense may have had an impact on responses."* (p. 222). But instead of leaving out this information, they choose to provide the participants with one specific type of task-irrelevant potentially biasing information.

While participants were provided with this information, they were kept in the dark about the following essential information. The authors chose not to define (or even rank) the verbal expressions of which they sought to study the perception, and claim that "this would have confounded any findings regarding perception accuracy". Similarly, participants did not receive guidance in how to map their perceived evidential strength on a 20-point scale from 'No support' to 'Conclusive'. Their choice not to define the latter scale numerically was additionally and similarly based on the claim "that this would have had a confounding effect [on] the findings of the experiment".

We do not understand the authors' claim that they can study the perception better by leaving undefined what was perceived and how to express this on yet another undefined scale, and by actively making it impossible *"to rely on some form of numeric reasoning"* (p. 222). This study calls for *"ensuring clear, effective and unambiguous communication when it comes to conveying the strength of evidence"* (p.227), but in the experiment does everything to avoid it.

The construction of the authors' own 20-point scale is inherently problematic. Participants were asked to express their perception of the verbal expressions on a column of 20 dots. Strength of evidence (LR) can vary from zero to infinity. All evidential strengths

supporting the defence proposition would be squeezed between the undefined zero and one points on the scale. It is equally unclear how to map the segment from one to infinity on (part of) a 20 point scale. This can provide the most obvious explanation for the mentioned *"similarity of perceptions at the higher end of the scale"* (p. 227). Ironically, we expect that the same effect would be found if participants were explicitly asked to map a quantitative expression of evidential strength on a 20-point scale.

A further issue is the study's 'between-subjects' design. Each participant was given only one verbal conclusion to be expressed on the authors' 20-point scale. This allowed the researchers to compare the expression of evidential strength between individuals and between the levels of the verbal scale. A 'mixed between-within' design would have been much better suited to study the distinction between the expressions of evidential strength. Having multiple expressions of different conclusions of each participant would have given valuable information on whether individual participants were able to distinguish verbal scale levels. The authors' Figures 3, 4, and 5 would likely have told a different story.

A confined scale such as the 20-point scale used in this study nicely fits on a piece of paper, but also reinforces the fallacious idea that it concerns a probability (of the prosecution's proposition being true), which is indeed confined between zero and one. The core problem is that the concept of strength of evidence is not sufficiently understood. This is true for both the participants in this study and for legal decision makers, who – although making decisions is their daily work – receive little to no education in even very basic decision theory.

We understand that in the adversarial system it is quite difficult to provide a lay jury with the appropriate knowledge to interpret expressions of evidential strength. It is nevertheless essential to do so. The results of the present study stress this once again. They also stress the need to include and define the full verbal scale, when such a scale is used to express the strength of evidence.

The main title of this study is "Understanding forensic expert evaluative evidence". We suggest that the relevant problem is not the perception of verbal expressions when their meaning is hidden, but the level of education and understanding of evidence evaluation and basic decision theory.

## References

[1] Arscott, E.; Morgan, R.; Meakin, G.; French, J., Understanding forensic expert evaluative evidence: A study of the perception of verbal expressions of the strength of evidence, Science & Justice 57 (2017) 221-227. (http://dx.doi.org/10.1016/j.scijus.2017.02.002)

[2] Association of Forensic Science Providers, Standards for the formulation of evaluative forensic science expert opinion, Science & Justice 49 (2009) 161–164. (http://dx.doi.org/10.1016/j.scijus.2009.07.004)

[3] Aitken, C.C.G.; Barrett, A.; Berger, C.E.H.; Biedermann, A.; Champod, C.; Hicks, T.N.; Lucena-Molina, J.; Lunt, L.; McDermott, S.; McKenna, L.; Nordgaard, A.; O'Donnell, G.; Rasmusson, B.; Sjerps, M.J.; Taroni, F.; Willis, S.M.; Zadora, G., ENFSI guideline for evaluative reporting in forensic science, 2015, European Network of Forensic Science Institutes (ENFSI). (http://enfsi.eu/wp-content/uploads/2016/09/m1_guideline.pdf)

[4] 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. (http://dx.doi.org/10.1037/lhb0000027)

[5] Fernbach, P.M.; Darlow, A.; Sloman, S.A., When good evidence goes bad: the weak evidence effect in judgment and decision-making, Cognition 119 (2011) 459-467. (http://dx.doi.org/10.1016/j.cognition.2011.01.013)