"Conclusions…": Signature and Handwriting Conclusion Terminology and Scales Panel Discussion

Abstract:

A current and global issue in our field is the topic of conclusion terminology and conclusion scales, particularly in respect of signature and handwriting conclusions. It is an important yet difficult topic to address because, while there is some commonality in the conclusion scales used in different geographical regions around the world, within a number of geographical regions there are multiple scales in use. It is for this very reason that it is also a topic in great need of discussion and there is a strong argument that we should attempt to reach a consensus (even if the result is that we agree to disagree).

This panel discussion is a collaboration of insights from numerous colleagues in our field in person, via Skype and in writing from private and government laboratories in geographical regions across the Americas, Australia, Asia, Africa, the Middle East and Europe.

Panel discussion contributor name:
R. Brent Ostrum

Laboratory currently employed with:
Canada Border Services Agency, Science and Engineering Directorate

Previous lab history:
Royal Canadian Mounted Police (National Police Services), Forensic Laboratory Regina

Short Bio:
Mr. Ostrum is a Senior Forensic Document Examiner in the Canada Border Services Agency's (CBSA) Science and Engineering Directorate in Ottawa, Ontario, Canada. Brent has worked as a forensic document examiner for around twenty-seven years being employed first by the Royal Canadian Mounted Police (obtaining his training there) and subsequently by the CBSA, in each agency's respective laboratory systems.

Brent is an A.B.F.D.E. Diplomate and the sitting chairman of Document Section of the Canadian Society of Forensic Science (CSFS). He is also a volunteer member of the Executive Council for Skill-Task Training Assessment & Research, Inc. (St^2ar) and has been active in standards and methods development for both forensic document examination (SWGDOC) and facial identification (FISWG). He has also served on the Program Committee for two 'International Workshop on Computational Forensics' (IWCF) meetings.

Brent has done case examinations resulting in testimony in various Canadian jurisdictions. His research interests include the assessment of examiner competency, method validation issues and the broad application of image and data processing to FDE problems. He is a student of logic, statistical inference, and evidence evaluation and, to further that study, he successfully completed the University of Lausanne course "Statistics and Evaluation of Forensic Evidence".

What conclusion terminology scale/s do you use? ("the scale/s")
Our lab presently uses a 7-level scale based on ASTM E1658/SWGDOC wording which endorses the expression of posterior odds statements about the propositions, including definitive conclusions of Identification and Elimination. In other words, we use a very 'traditional' conclusion scale.

Have you or your laboratory changed conclusion terminology scales over the years? If so, why?
Yes, this has happened a few times in my career. The reasons for the changes vary by instance but include attempts to 'improve' the wording (generally to the subjective liking of section members),

for better internal consistency in the 'explanation' of different levels (again, to the subjective liking of section members), or to fulfil some management decision regarding such things (to the subjective liking of persons in authority). Rarely, if ever, have the changes been made based upon empirical research indicating that a particular conclusion scale or wording was more appropriate, better supported or effective.

Are you comfortable using the scale/s or are you confined to using that scale/s because of lab requirements?
This is an interesting question. To some degree, I am 'confined' to it primarily because it is our present lab policy. At the same time, our policy is reasonably flexible insofar as examiners are permitted to express conclusions using other wording should it be warranted. In those instances, the examiner and peer reviewing examiner (and often the section manager) discuss the situation and decided on the best course of action.

In addition, I am comfortable using the scale described above even though it is not my preferred choice. As should be apparent from my following answers, I think there are better options and I am an advocate for the use of a logical approach to evidence evaluation which lends itself best to opinions expressed in the form of a likelihood ratio, or LR, (ie. statements about the probability of the evidence given competing propositions).

However, it is very important to understand that preference for a logical approach does NOT mean that an examiner cannot use a scale derived from E1658, i.e., based on posterior odds of the propositions. To the contrary, an examiner may do so if 1) they are able to address the logical issues in moving from the LR to posterior odds, and 2) they address the issue of the decision heuristic used to declare definitive opinions. The specific nature of those issues, and how they should be addressed (or avoided), are made completely clear in the logical approach. I will expand on this in later answers.

Do you consider the scale (or, if multiple scales are in use, which one) to be the most appropriate of the various scales in use around the world in today's professional climate? If so, why? If not, which scale do you consider to be most appropriate and why?

The scale used in our lab (and in most other labs) is not the most appropriate option in my opinion. At least, it is not without careful and thorough explanation of the assumptions made by the examiner to justify extension of their opinion beyond the evidence to encompass the proposition, as per my earlier comments).

First I will address the primary reasons why I believe our scales, and similar 'traditional' scales are inappropriate.

In my experience, most examiners don't understand the issues that arise when expressing 'traditional' conclusions that take the form of posterior odds. As a result, examiners routinely make inferences and express conclusions that, quite simply, extend far beyond what is warranted given the expert's role in the overall decision-making process. And they do so in blissful ignorance of the situation.

Keeping things very brief, extension of our conclusions in this manner means that the examiner must make a number of (generally debatable and certainly personal) choices and decisions about the prior odds of the propositions. If they are able to do so while explaining those choices clearly to the recipient of their conclusion, then it would be possible to express conclusions in the form of posterior odds without over-stepping the bounds of logic and propriety. Expressing definitive

conclusions of identification or elimination goes a bit further still. For those conclusions the examiner must also explain how and why they feel any residual potential for error in the decision can be dismissed or ignored. But, if they can do that, then even those conclusions will be logically sound and acceptable.

Thus, in a certain sense there is no real problem with any conclusion scales if the examiner can reconcile the issues of logic hidden in our present approach to evaluation and conclusion wording.

In the practical sense, however, that is a non-trivial task. I believe that once an examiner understands the limits and issues that exist they are faced with two simple choices. The easiest and most sensible is to limit oneself to statements about the evidence given the propositions of interest. In other words, some form of likelihood-ratio for the evidence. The second, and far more difficult, is to extend beyond this while explaining why and how their personal (and not scientifically derived) beliefs should be acceptable to the trier/client.

Now I will turn to the conclusion 'scale' I consider to be the most appropriate. The best option, in my opinion, is a (coherent) logical approach to evidence evaluation. Please note that, strictly speaking, I am not referring only to our 'conclusion scale' or the wording of our opinions. Instead, I am talking about the manner in which we evaluate evidence in its entirety.

I believe that we should not discuss the 'conclusion scale' or wording used to express the final result without fully understanding the evaluation procedure that leads to the conclusion. These two elements are inseparable and should not be addressed in isolation from one another.

Ultimately the use of a coherent logical approach favours opinions expressed in the form of a likelihood ratio, or LR, (ie. statements about the probability of the evidence given competing propositions) because that is a direct reflection of the evaluation process itself. In other words, the conclusion flows directly and logically out of the evaluation process.

Our goal, as experts, should be to help our clients and the courts to understand the evidence that falls in our domain, in terms relevant to the matter at hand. It is the role of the client or court to decide about what really happened in any given situation, using whatever method suits them. More specifically, it is the role of the trier-of-fact, the decision-maker, to determine which of the various competing propositions is best supported by the evidence (all of it) and whether that degree of support fulfils whatever legal criteria applies in that instance.

Do you feel passionately about (for or against) any other conclusion terminology scales currently in use around the world?

Yes and no. I have outlined most of the general issues above.

Traditional scales speaking to posterior odds of propositions are, as a general rule, logically incoherent and unsupportable without a number of dubious assumptions or assertions. As such, I feel they are inappropriate and should not be used by FDEs, or other forensic scientists for that matter. In that regard I do feel passionately that our discipline would be addressing those issues.

I also feel strongly that the best conclusion wording derives from a correct and proper evaluation focused on the likelihood-ratio both for the assessment of the evidence and expression of the result.

However, this question asks about "*conclusion terminology scales currently in use around the world".*

While I advocate the use of LR wording in conclusions, there are issues with many of the existing scales based upon this approach (and there are several). First, existing 'scales' based on the LR

approach can and do take a number of different forms. That isn't necessarily a problem but it does complicate matters when trying to explain what is going on – or when determining how the different LR scales relate to one another. A few, such as those used by the NFI or SKL, are well-structured and appropriately formed LR scales. On the negative side, they are complicated and have a large number of levels in the scale. I believe that the number of levels which are appropriate and justified is open to debate. It will take considerable research to fine-tune our knowledge to shown precisely how many levels are warranted and justified. At this time, I would argue that any such scale might be valid for up to 4 levels per component of the LR producing a maximum number of 'final conclusions' on the order of 7, as shown in the following table (using wording of an 'indirect' nature to address the support provided by the evidence for each of two propositions, X and Y):

	Very strong support for X	***Strong support for X***	***Moderate support for X***	***Low support for X***
Very strong support for Y	**Inconclusive** – 'equal' support for both X and Y	Moderate support for Y over X	Strong support for Y over X	Very strong support for Y over X
Strong support for Y	Moderate support for X over Y	**Inconclusive**	Moderate support for Y over X	Strong support for Y over X
Moderate support for Y	Strong support for X over Y	Moderate support for X over Y	**Inconclusive**	Moderate support for Y over X
Low support for Y	Very strong support for X over Y	Strong support for X over Y	Moderate support for X over Y	**Inconclusive**

In my opinion, there is little reason to believe that examiners can reliably or accurately distinguish a greater number of levels. Research will have to guide us on that. I should probably note that this issue exists with ALL the present scales being used or proposed, including those based on ASTM/SWGDOC wording.

I characterize some other scales as being quasi-LR in nature. These include the scale proposed by the UK AFSP and the Australian DocSAG. The big problem with these scales is the one-sided nature of the wording, despite there being generally correct discussion of the approach in accompanying documentation. As the above table clearly shows, any level of support for a single proposition may still prove to be 'inconclusive' if the support for the alternative proposition happens to be at the same level. What matters in all of this is the **differential** level of support provided, not the support for any one proposition taken alone. Any scale that speaks to support for one proposition without concern for any specific alternative is incomplete and insufficient.

So I have concerns with most of the present approaches to using/expressing conclusions in the form of the LR.

Of the various options under consideration I prefer, and this is just a personal preference, the use of wording that speaks to the support provided by the evidence (for each of the propositions), rather than the probability of the evidence (given the propositions). I think of the former as speaking to the probability 'indirectly', whereas the latter does so 'directly'. Until there comes a time when examiners, as a group, are more comfortable expressing their personal beliefs in terms of numeric probabilities (whether subjectively or empirically derived) I believe that 'indirect' wording is a better choice.

In the end I feel that whatever wording is used **MUST:** 1) reflect the actual examination and all of the limitations that exist in that process, and 2) be logically coherent and sound. More specifically, and using terminology developed by far brighter minds than mine, there are four **essential** requirements that must be met by any evaluation and reporting scheme.

These requirements are: Balance[1], Logic[2], Robustness[3] and Transparency[4].

Do you believe that there should be international unification on conclusion terminology and the scale used in signature and handwriting conclusions? Why?

Unification is not absolutely essential but it would certainly be beneficial. In the first place, examiners are doing the 'same thing' regardless of where we happen to live and work. There is no good reason to use different terminology or scales simply because the work is done in one location versus another. Science and logic do not change by geography. Translating our results into layperson-speak need not vary by geography either. For what it is worth, I would suggest that even if the method varies somewhat by geography there would be no reason to use different terminology.

A unified approach would help to minimize confusion about which opinions are appropriate and supportable, and which are not. Courts are increasingly concerned about forensic evidence and expert testimony and they routinely look to other jurisdictions to determine what is appropriate for this type of evidence. At an international level legal rulings on 'science' and expertise seem to be converging with many of the same elements arising, regardless of what court is involved.

Unification would undoubtedly improve communication between examiners and with clients/end users, as well as promote better logic, improve transparency, and so on.

Some have argued that different court/justice systems place different demands or restrictions upon what an examiner can or should say during testimony. That certainly may affect how some examiners report their opinions. At the same time, I find it difficult to believe that any court interested in true justice would force an examiner to give a conclusion that cannot be supported or sustained under challenge.

In addition, the logical approach to evidence evaluation does not 'stop' or preclude any form of wording for our opinions. It simply helps us understand what is needed in order to justify such wording. If justified, there is no problem.

Other matters of relevance and importance relating to this topic that you would like to discuss.

---

[1] 'Balance' means that the evidence/findings should be evaluated given at least one pair of competing propositions; ideally with the first proposition based upon one party's account of the events and the latter based upon an alternative account.

[2] 'Logic' means that the evaluation process must be one that speaks first to the probability of the evidence/findings given the propositions (plus relevant background information), and not the probability of the propositions given the evidence/findings (plus background information). This is essential to ensure there is no inappropriate or unjustified transposition of the conditional since proper or correct transposition of the conditional requires information generally not within the scope of the examiner.

[3] 'Robustness' means simply that the evaluation process must be capable of sustaining scrutiny or review by other experts through review or cross-examination. It should be based upon sound knowledge and experience of the evidence type including the use, when available, of pertinent databases, published data or ad hoc case based experimentation. In other words, 'robustness' refers to the scientist's ability to explain the grounds for their opinion based upon their degree of understanding of the particular trace type and it's probability of occurrence in the relevant 'population' relating to each of the competing propositions.

[4] 'Transparency' applies to all facets of the examination and evaluation and means that the entire process should be demonstrable and recorded so as to permit proper review and assessment. Worknotes should clarify all relevant aspects of the evidence including the interpretation and evaluation of that evidence in terms of the competing propositions. The report should be written in way that is suitable for a varied audience (i.e. participants in the justice system).

I am the first to admit that the adoption of any LR-based approach or system faces many hurdles.

First, there is a serious lack of knowledge and understanding by examiners. I find this troubling given that we profess to be experts in all matters relating to our discipline. At the same time, I acknowledge that it isn't a trivial or easy thing to understand or to adopt. But it isn't that hard either. The bigger issue is that people need to understand why this topic matters so that we can have meaningful discussions about how best to proceed. A great example is today's session and I applaud the ASQDE for putting this on the agenda.

Second, there is a lack of knowledge and understanding by the courts (and lawyers). The level varies, of course, by court and jurisdiction but generally speaking it is not much better than the level of knowledge for FDEs. Put simply, most clients and courts do not understand the logical approach very well at all. As a result, there has been some concern about whether they will 'accept' this type of testimony. I believe this concern to be overblown. First, our traditional approach is hardly simpler or easier to understand. It is simply familiar and the flaws and issues inherent to it are unknown. In that regard, I am reminded of a quote by a learned participant on this panel. He put it nicely. "A logically incorrect conclusion that is 'understood' is no alternative to a logically correct conclusion which needs explanation."

In addition, there are signs that this may be changing as various review bodies are developing standards and guidelines that speak to these issues. Even in the US and Canada there are activities aiming to address this issue. I suggest there is a slow but definite movement towards the adoption of this approach regardless of the desires of practitioners or the courts.

Third, the logical approach permits a number of scales (various types of wording are possible as well as multiple options for levels and structure). As a result, work is still needed to standardize those aspects in a meaningful way.

Lastly, there remains a huge need for additional research into related topics such as bias and objective quantification. That is the case no matter what approach or conclusion scale is being considered. However, in my opinion, the LR or logical approach provides the best option to address those elements as and when the research is done.

In spite of any outstanding challenges or issues the logical approach is rigorous in the application of logic in a coherent manner. It is firmly focused on the aspect of evidence to which we (as examiners) can safely and correctly speak.