Review of 'Treatment of Inconclusives in the AFTE Range of Conclusions' For Law, Probability, and Risk

Note to the Reviewer(s)

Thank you very much for your extremely helpful comments. This paper is much better for your comments, and we have reworked portions of the paper to include the additional suggested studies, which makes our overall conclusions more useful and pointed. We have attached a PDF file showing the changes, with the exception of changes made in tables and equations - these have obviously changed, but the change tracking software is less than helpful with these structures. You can see that the figures showing data and the portions of the paper dealing explicitly with error rate studies are the areas which have changed the most in response to your comments and suggestions.

Review 1

This manuscript provides a systematic and interesting review of several experimental studies conducted to measure error rates in firearm examination. Error rate in the pattern matching disciplines of forensic science is an active and important subject of research. The article thoroughly discusses the various ways that error rates may be calculated and lays out the results from the studies clearly with useful illustrations and tables. Out of this analysis, the article adds an important observation about an asymmetry associated with inconclusive examiner opinions, that they are significantly associated more with different source pairs than same source pairs. This observation seems to point to an bias in the subjective decisions by firearm examiners. In my opinion, the article will be acceptable for publication in Law, Probability, and Risk after the authors have considered and responded to five issues described just below. In addition, I provide several other optional suggestions and questions for the authors to consider before publication.

Main points

1. Section 2.3, final two paragraphs on page 7
   The two paragraphs below clearly seem to contradict one another and must be reworded: "Complementary to accuracy rates, we define two error rates of a classification method: The false positive rate (FPR) - or failed elimination rate - is the probability that an examination of different source evidence does not result in an elimination. Similarly, the false negative rate (FNR) - or missed identification rate - is the probability that an examination of same-source evidence does not result in an identification. Note that we are keeping the definition of error rates a bit vague on purpose. As discussed in Section 2.2, there are multiple ways to define an error, but the concepts of FPR and FNR hold for all of the options of treating inconclusive decisions outlined in Table 3."
   If the FPR is the probability that an examination of different source evidence does not result in an elimination, then inconclusives always count as errors. A similar argument holds for FNR. Therefore, the concepts of FPR and FNR as defined above are only consistent with option 3 in Table 3 and not with all the methods of treating inconclusives as stated in the second paragraph.

   - Thank you for pointing out this inconsistency - we tried to define error rates conceptually first before getting into the more contentious areas of what constitutes an error. We have rephrased the first paragraph to:
     The false positive rate *(FPR) – or failed elimination rate – is the probability that an examination of different source evidence* ***results in an error****. Similarly, the* false negative rate *(FNR) – or*

missed identification rate – is the probability that an examination of same-source evidence ***results in an error****.*

2. Page 19, first paragraph below figure 8
   The following two sentences seem inconsistent: "There is an order of magnitude difference between the probability of same-source evidence given an elimination and the probability of different-source evidence given an identification. Figure 8 provides a visual demonstration of this discrepancy: unlike in Figure 1, there is a difference in the density of points in the inconclusive sections of the chart." The first sentence refers to identifications/eliminations whereas the second sentence refers to inconclusives. In addition, Fig. 8 does not show the order of magnitude difference between the probability of same- source evidence given an elimination and the probability of different-source evidence given an identification because there are four dots in the one field and nine dots in the other field. Taking these sentences together, it is not clear whether the disturbing realization in the previous sentence refers to eliminations/identifications or to inconclusives. Since this is one of the most important observations of the manuscript, it is important to state the observation clearly.

   - We have revised this section as part of the incorporation of a number of papers that approach error rates from a European perspective. We have additionally added references to the equations and the relevant figures, and refer to both. In addition, we have added an explanation of the sample size issues - when counts are not observed for e.g. same-source eliminations or different-source identifications, we have a hard time properly estimating the equivalent rates.

3. Page 19, paragraph 3, lines 2-4
   The following statement is false: "We see that in the three studies with a large enough sample size, the point estimate of the probability of a false identification is higher than the probability of a false elimination." Figure 7 clearly shows that the probability of a false identification and the probability of a false elimination are both equal to zero for the Keisler data.

   - This sentence was revised as part of the inclusion of additional European studies. The original point was to suggest that Keisler was not sufficiently large (by the criteria explained above), but we have added language to clarify this criteria and have also rephrased this explanation to hopefully be clearer.

4. The manuscript does not provide any basis for the observation, "... in many studies, there are more same-source comparisons than what would be expected in case work." They must eliminate the observation or provide evidence for the relative amounts of same source and different source comparisons found in case work.

   - This was changed and reduced a bit to
     For example, in many studies, the study is designed such that it is possible to estimate false elimination rates more precisely than false identification rates. In fact, in some common study designs, it is only possible to estimate the false elimination error rate; the rate of false identifications cannot be estimated at all.
     We believe this statement is fully supported by the available evidence, where the previous statement was partially drawn from conversations with examiners that cannot be easily cited.

5. Page 29, table 11
   The numbers 0.2414 and 0.7586 in the right-hand column seem incorrect. I would have expected to see 0.1944 (=70/360) and 0.8056 (=290/360).

   - Thanks for catching that, we must have checked the numbers 10x or more, but errors still sneak in! We have re-checked the numbers and reformatted the tables slightly to emphasize the difference in the calculations of source-specific and decision-specific conditional probabilities.

Additional Issues

In addition, the points below are for optional consideration by the authors before publication

- Abstract: The abstract could be improved by stating any important observations there, for example that inconclusive examiner opinions are significantly associated more with different source pairs than same source pairs.
  - We have added some additional information to the abstract, but have described our findings with an eye to minimizing inflammatory language so as not to bias any potential readers before they understand our assumptions and logical foundations.
- Section 2.4, page 10, lines 1-2:
  Upon first reading, the clause "however, the results of designed studies are used to give information tailored to the situation in a particular case" gave me the mistaken impression that the studies are designed to give information about case work. This impression could be mitigated by changing the clause to read "however, the results of designed studies are used here to give information for the different decision scenarios," or by a comparable change.
  - changed. Thanks!
- Page 15, section 3.4, last sentence:
  The following sentence seems overstated: "Regardless of the explanation, it is apparent that when presented with a different source comparison, examiners make an elimination much less frequently than an identification when presented with a same-source comparison." Given the elimination data shown in in Fig. 5 where the differences are less than an order of magnitude, I would change "much" to "significantly".
  - I've removed the additional word entirely - "significantly" has meaning in statistics that is not completely supported here (because most studies are not big enough to draw a statistically significant conclusion), but I wouldn't want to use "much" for the same reason.
- Page 16, last two paragraphs of section 3.5:
  - The term "target probability" is not defined in the last paragraph. Upon review it becomes clear that the term is exchangeable with the term "expected proportion" in the preceding paragraph. I would change "expected proportion" to "expected proportion (target probability)".
  - changed, thanks for the help with text clarity
- Page 17, line 8 below Table 6:
  To improve clarity, I would change "This implies that we can calculate the"target" probability for each decision" to "This implies that we can calculate the"target" probability for each decision for same source pairs".
  - changed (slight variation)
- Page 18, figure caption 7:
  The statement that the Baldwin study is the only study with a large number of participants and evaluations is not consistent with a previous statement that the Keisler and Duez studies were also large. I would state rather that the Baldwin study is the largest of the three studies shown in the figure.
  - I'm unable to find previous statements implying that Keisler and Duez are large studies.
  - We've added language to explicitly clarify what constitutes a sufficiently "large" study, and have also added a couple of EU-focused studies that also have a relatively large number of comparisons. Additional text compares the two sets of studies (US/CA vs. EU) and points out that the conclusions of bias do not hold uniformly for both sets of studies.
- Page 19, line 3 from the bottom:
  Change "criteria" to "criterion", because the intended meaning is singular.
  - done
- Page 20, last sentence:
  I did not understand the sentence: "We also suggest that during cross-examination after the testimony was presented, the defense ask about the decision specific probability of an error (relative to the

examiner's decision), which will provide specific information which is relevant to the testimony at hand." Please reword.

- Changed to: *Specifically, in court, we suggest that when the admissibility of an examiner's testimony is assessed, the examiner is asked to state the lab policy on making eliminations and the rate at which the examiner makes inconclusive decisions, along with any relevant error rates specific to the lab and to themselves. We also suggest that during cross-examination after the testimony was presented, the defense ask about the decision-specific probability of an error (relative to the examiner's decision). This information will provide error rates which are specific and relevant to the presented testimony.*

Review Addendum

The authors have not included in their analysis several published studies from European laboratories that dealt with error rate testing.Perhaps these studies were not included because the emphasis of the manuscript is on the AFTE theory of identification, rather than the ENFSI method of estimating likelihood ratios. It would be beneficial to check the results of these studies to make sure that the data do not change the principal observations. It would also add to the usefulness of the manuscript to cite the articles. They are as follows:

- Pauw-Vugts, P., et al., FAID2009: Proficiency Test and Workshop. AFTE Journal, 2013. 45(2 -Spring): p. 115-127.
- Mattijssen, E., et al., Validity and reliability of forensic firearm examiners. Forensic Sci Int, 2020. 307, 110112.

Finally, more black box test data using VCM have been published by the group that wrote the Duez et al. article. These data would add to or supersede the data shown in Table 9 of the manuscript. See

- Lilien,R., Firearm Forensics Black-Box Studies for Examiners and Algorithms using Measured 3D Surface Topographies, NIJ Report 254338. 2019, Office of Justice Programs: Washington DC. p. 38.
- Chapnick, Chad et al., Results of the 3D Virtual Comparison Microscopy Error Rate (VCMER) Study for firearm forensics, J. Forensic Sci., 2020. DOI: 10.1111/1556-4029.14602

These studies have been incorporated into the paper and the differences between the European studies and the US/CA centric studies have been explicitly discussed. We don't believe that the inclusion of these studies changes the principle observations of bias - rather, they help to show that this bias does not necessarily exist under all conditions, but that we do not currently have enough research to specifically determine which factors (legal system, examiner training, reporting scales, physical microscopy vs. virtual microscopy, etc.) might be most important in explaining the difference