Legal probabilism

Marcello Di Bello – ASU - November 14, 2024

Notes on Eyewitness Confidence and Accuracy

The claim

1. When an eyewitness makes an identification (”yes, it was him!”), should we believe them? That eyewitness evidence is prone to error is well-known—all evidence, in the end, is prone to error. A specific question we tackle here is whether someone’s confidence is an indicator of accuracy. That is, if an eyewitness makes an identification with great confidence, should we then conclude that the eyewitness is very likely correct?

2. Wixten and Wells in their 2017 paper ‘The relationship between Eyewitness Confidence and Identification Accuracy’ answer affirmatively, at least provided the identification was made in lineups under *pristine conditions* (more below). By contrast, the legal system and previous literature in psychology hold otherwise. For example, here is an excerpt from a brief by the American Psychological Association (cited in the paper):

   > “In one article reporting results from an empirical study, researchers found that among witnesses who made positive identifications, as many as 40 percent were mistaken, yet they declared themselves to be 90 percent to 100 percent confident in the accuracy of their identifications. . . . This confirms that many witnesses are overconfident in their identification decisions.” (p. 12)

   So, that eyewitness confidence is a good indicator of accuracy is far from commonplace!

3. What are pristine conditions? Eyewitness identifications take place during police lineups. Lineups comprise a suspect and a few other filler individuals. A witness to the crime is asked to pick one individual from the lineup if they recognize them as the perpetrator. Lineups are pristine when (p. 20):

   (a) the lineup includes only one suspect, all other individuals are fillers

   (b) the suspect should not stand out compared to the fillers

   (c) the officer who administers the lineup cautions that the perpetrator might not be present

   (d) neither the witness nor the officer know who the suspect is (double blinding)

   (e) a confidence statement is collected at the time of the identification

4. So, in summary, the claim of the paper is this:

   > “if the pristine conditions ... are followed, then a low-confidence ID implies low accuracy, and a high-confidence ID implies high accuracy.” (p. 20)

   But how should accuracy be measured? This is the next question we tackle.

MEASURING THE CONFIDENCE/ACCURACY RELATIONSHIP

5. The relationship between confidence and accuracy can be measured in different ways. Perhaps surprisingly, the same data can tell very different stories about this relationship depending on the measure used. Here are three key measures:

   Correlation Correct responses are coded as 1's and incorrect ones are coded as 0's. These outcomes are plotted by confidence levels (say between 0 and 5), as in Figure A1 p. 56.

   Calibration Calibration measures accuracy as the percentage of correct identifications out of all identifications:

   $$100 \times \frac{\#correctIDs}{\#IDs},$$

   for each confidence level. Under perfect calibration, if witnesses are 80% confident in their IDs, they would be correct 80% of the time.

   Confidence-Accuracy Characteristic (CAC) This measure is similar to calibration, with the difference that the denominator only includes IDs of suspects (more below):

   $$100 \times \frac{\#correctIDs}{\#suspectIDs},$$

   again for each confidence level. Since in lab experiments there is no clear suspect—unlike real world police lineups in which one individual is the suspect—this measure can be computed by designating one individual in the lineup as the innocent suspect or by dividing by the lineup size. (p. 25). These two approaches are roughly equivalent.

6. Wixten and Wells hold that the CAC measure is the best. Calibration is a more intuitive measure than correlation, but one limitation of calibration is that incorrect IDs apply to both suspects and fillers. However, only incorrect IDs of suspects are problematic: if an innocent suspect is identified in a lineup, they may face trial, but if an innocent filler is identified, they will not go to trial. This limitation makes CAC preferable:

   > "if the eyewitness picked a filler, we already know that the witness did not pick the perpetrator. So, the forensically relevant question is this: Given that the eyewitness picked the suspect with a particular level of confidence, how likely is it that the suspect is guilty? The answer to that question is provided by a CAC plot in which the dependent measure is suspect-ID accuracy. " (p. 24)

7. These measures are easily applicable in laboratory studies so long as the ground-truth is know. What did the available laboratory studies find? Correlations were initially found to be positive but low. Later studies that focused only on positive identifications (so-called "choosers") showed stronger correlations between confidence and accuracy. Calibration studies made this point clearer:

   > "Calibration studies typically find a strong relationship between confidence and accuracy when (a) the analysis is limited to choosers, (b) the witnesses are adults, (c) the lineups are fair, and (d) the confidence ratings are taken immediately after the ID is made" (p. 23)

But calibration studies also showed overconfidence. That is, witnesses are who extremely confident (say between 90% and 100% confident), do make mistakes more than 10% of the time. Finally, when data of existing studies—using correlation or calibration as measured—are reanalyzed, a near perfect correlation emerge between accuracy and confidence (pp. 26-37). See Fig. 5, p. 37. This is the technical result by Wixten and Wells.

FIELD STUDIES

8. Laboratory studies work well because the ground truth is known, but they may not reproduce conditions in the real world:

   > "The advantage of a mock-crime study such as the ones considered above is that the experimenter knows if a suspect ID is correct or incorrect, thereby allowing a direct computation of suspect-ID accuracy. In a police department field study, by contrast, it is not known if a suspect ID is correct or incorrect. Thus, although one can measure how often high-confidence and low-confidence IDs are made to suspects and fillers, a direct calculation of suspect-ID accuracy as a function of confidence is not possible." (p. 30)

9. However, field studies of police lineups also suggest a strong correlation between confidence and accuracy. For example, in the Hennepin County field study:

   > "Of 175 choosers in this study, 96 (55%) made jump-out IDs (=with near certainty confidence). Remarkably, 99% of these IDs were made to suspects, not fillers, which is to say that only one of the 96 jump-out IDs was made to a filler. ... there were 5 times as many fillers as suspects in any given lineup, so random responding for jump-out IDs would result in 26 $\times(5/6) \approx 22$ filler IDs (yet only one was actually observed) and only about $26 \times (1/6) \approx 4$ suspect IDs (yet 25 were actually observed). Thus, the number of suspect IDs made with high confidence in this study was far greater than would be expected by chance. It is possible that the lineups in the Hennepin County study were not fair lineups. But if they were fair lineups (as they were designed to be), it is hard to come up with a logical explanation for these results without assuming that high-confidence accuracy was close to perfect." (p. 30)

BASE RATES

10. A further problem is the role of base rates of perpetrator-present lineups. How often is the actual perpetrator present in lineups? Laboratory studies usually assume a 50% rate, but we don't know whether this figure is the same as in real police lineups.

11. Why does the base rate matter? As Figure 8 shows (p. 42), accuracy decreases dramatically as the base rates goes down below 35%, even for witnesses who are very confident in their identification. This should not be surprising. As Bayes's theorem tells us (see more detailed discussion in Appendix C):

    > "Specifically, the probability that a suspect is guilty given that the witness identified that suspect is a function of both the diagnostic value of the evidence and the base-rate probability that a lineup's suspect is guilty." (p. 60)

12. Can we know what the base rates are in police lineups? This question is hard to answer. But the good news is that

> " using pristine identification procedures, the laboratory data shown in Figure 8 suggest that, at a base rate of only 35%, confidence is highly predictive of suspect-ID accuracy, and high-confidence IDs are still quite accurate (about 90% in Fig. 8), whereas low-confidence IDs, despite having probative value, are highly error prone." (p. 43)

13. The importance of base rate should also make them a more central focus of investigation:

> "Moreover, base rates likely differ from jurisdiction to jurisdiction, which means that some may fall well below the 35% estimate in Houston. Thus, conceptualizing the base rate as a system variable—and taking concrete steps to increase it—seems like a prudent strategy for law enforcement to consider." (p. 43)

SIDE REMARK ON LEGAL PROBABILISM

14. No matter the details, the research on eyewitness evidence shows that probability is crucial to study, analyze and improve the accuracy of police lineups. Probability gives us a language to talk about eyewitness accuracy that we would not have without it. See also how base rates affect the assessment of eyewitness accuracy, an insight made precise by Bayes's theorem.

15. More generally, the research on eyewitness evidence suggests a third use of probability theory besides the other two we have seen so far in the course:

    (i) weigh and assess evidence,
    (ii) model decision-making, and
    (iii) analyze the performance of practices and procedures of the justice system such as police lineups and eyewitness IDs.