Legal system v. eyewitness

Colloff, Melissa; Wilson, Brent; Flowe, Heather

DOI:
10.1016/j.jarmac.2021.04.001

Legal System v. Eyewitness: The Jury is Still Out on Who is Better Able to Reduce Eyewitness Error (Variance)

Melissa F. Colloff[a], Brent M. Wilson[b], & Heather D. Flowe[a]

[a] School of Psychology, University of Birmingham, UK
[b] Department of Psychology, University of California, San Diego

Author Note

Correspondence concerning this article should be addressed to Melissa F. Colloff, School of Psychology, University of Birmingham, Birmingham, UK, B15 2TT. Email: M.Colloff@bham.ac.uk

Word Count: 2914

Keywords: eyewitness identification, error variance, correlated memory signals, reliability, discriminability

Author contributions: All authors discussed the content, MFC and BMW drafted the paper and all authors provided critical revisions.

Brewer and colleagues have introduced and researched a *confidence identification procedure* wherein witnesses rate how confident they are that each person in the lineup is the culprit (Brewer et al., 2012, 2020; Sauer et al., 2008, 2012). The goal is to determine the likely guilt or innocence of the suspect using confidence ratings, which ostensibly measure the degree of match between each lineup face and the witness's memory of the culprit. The approach seeks to increase the information that can be collected from witnesses over the categorical decision (suspect ID, filler ID, lineup rejection) witnesses make in a traditional lineup procedure. The confidence procedure also seeks to leave the decision about the suspect's guilt in the hands of someone other than the witness. It is envisaged that suspicion regarding the guilt of the suspect would rise or fall with the police investigators' interpretation of the witness's confidence ratings and that fact finders would weigh the confidence evidence in their determination of the defendant's guilt (see also Gepshtein et al., 2020). We agree with Brewer and Doyle (2021) that there are intuitive benefits of maximizing the information that can be collected from eyewitnesses. Here, we outline two key steps that we deem to be critical before the procedure is considered for implementation.

1. Determine how the confidence procedure influences witness discriminability compared to the traditional procedure

Brewer and colleagues (2020) have demonstrated that the pattern of confidence ratings across the lineup members can be used to predict the probability that the suspect is the culprit using regression analyses. This analysis is conceptually similar to calibration or confidence-accuracy-characteristic analyses, where objective performance (i.e., probability of suspect guilt) is plotted against subjective confidence ratings (see Brewer & Wells, 2006; Mickes, 2015). These analyses measure *reliability*, which refers to the probability that a judgment made with a certain level of confidence is correct. This information is undoubtedly useful for legal decision makers, such as jurors and police officers (Mickes, 2015). A reliability analysis can help to answer the question: given the witness's decision (or pattern of confidence ratings), how much weight should legal decision makers attach to the witness's judgements? The confidence procedure appears to be a promising way to collect information to answer this question. As noted by Brewer and Doyle (2021), the confidence procedure ostensibly permits more detailed analysis of probability of suspect guilt using the pattern of confidence ratings compared to a traditional lineup, since there is no longer just a single rating to the person who is identified. Moreover, the confidence procedure always collects a confidence rating to the suspect, and a judgment about the suspect is more informative of likely guilt than a judgement about a filler or a lineup rejection (e.g., Starns et al., 2021).

An unanswered question, however, is whether policymakers should choose to employ the confidence procedure instead of the traditional procedure that it would replace (e.g., it is typical in the US for witnesses to see a 6-person simultaneous lineup and provide a confidence judgement). Unlike Brewer et al.'s (2020) regression analysis, which answers a question of interest to judges and jurors, a receiver operating characteristic (ROC) analysis answers the policymaker's question (Mickes, 2015). ROC analysis does not measure the posterior probability of guilt (i.e., the probability of guilt given a specific witness's pattern of confidence ratings) but instead measures the collective ability of witnesses to discriminate between innocent and guilty suspects (Macmillan & Creelman, 2005; Mickes et al., 2012). ROC curves are constructed by plotting hit rates of guilty suspects in target-present lineups against false alarm rates of innocent suspects in target-absent lineups, across different levels of response bias, typically measured by confidence (Mickes et al., 2012).[1] The higher the ROC—and the larger the area under the curve (AUC)—the better witnesses are at sorting innocent and guilty suspects into their correct categories. A procedure that yields a higher ROC results in a higher hit rate of guilty suspects and a lower false alarm rate of innocent suspects than a procedure that falls on a lower ROC. This is why it is sensible to base policymaking decisions about which lineup procedure to employ on the procedure that yields the highest ROC (Mickes, 2015; National Research Council, 2014).

An ROC analysis of the confidence procedure is important to conduct because it provides different information than a reliability analysis. Counterintuitively, two procedures that have comparable reliability can nevertheless differ dramatically in how well they correctly sort innocent and guilty suspects. For example, both simultaneous and sequential lineups seem to be equally reliable in that high-confidence suspect identifications result in similarly high probabilities of guilt, yet the simultaneous lineup typically has a higher ROC (e.g., see Colloff et al., in press; Mickes, 2015; Seale-Carlisle et al., 2019). The implication of this ROC advantage is that the simultaneous lineup would yield a larger number of reliable, high-confidence suspect identifications than the sequential lineup.

[1] Note that other researchers have recently devised their own ROC-like alternative that includes filler and reject decisions in the ROC (e.g., Smith, Yang, & Wells, 2020). In this comment, we are referring specifically to the use of ROC analysis in the traditional way that it has been used for decades in the basic scientific literature, directly tethered to a longstanding model of decision-making—signal detection theory (e.g., Mickes et al., 2012; Wixted, 2020.

Future work should compare the confidence procedure that Brewer and Doyle (2021) envisage being employed in practice (e.g., a sequential confidence procedure) against the traditional lineup procedure that it is set to replace. This work should jointly consider performance in target-present and target-absent lineups, and should include ROC analysis. Such work might find that the confidence procedure yields a higher ROC than the traditional procedure. For example, Brewer and colleagues suggest that the confidence procedure might reduce the influence of system and witness factors that impact the decision criterion set by a witness (Brewer et al., 2020; Brewer & Doyle, 2021). In that case, it is possible that the confidence procedure reduces criterion variability that can pull down the ROC (e.g., Mickes et al., 2017). Alternatively, both procedures could yield the same ROC. An outcome like that may occur if witnesses use the same strategy for making decisions and ratings on both a confidence and a traditional lineup. Finally, the traditional procedure could yield a higher ROC than the confidence procedure. An outcome like that might occur if the confidence procedure encourages witnesses to consider each face in isolation, rather than capitalizing on useful information that can be gleaned from comparison across similar-looking faces (e.g., Wixted & Mickes, 2014; Wixted et al., 2018). We explain this in detail, next.

Taking advantage of correlated memory signals

Memory signals generated by lineup fillers and the suspect are likely to be correlated because fillers share features with the suspect as the lineup members all match the witness' description of the culprit (Wixted et al., 2018). Consider two hypothetical eyewitnesses, Kat and Evan, who each see a 20-year-old White male committing a crime. Kat forms a strong memory of the perpetrator. When Kat views the lineup, the perpetrator generates a strong memory signal because he matches Kat's strong memory, but the other fillers also generate somewhat strong memory signals because they are also 20-year-old White males. Evan, however, forms a weak memory of the perpetrator. When Evan views the lineup, the perpetrator does not generate a strong memory signal because Evan does not have a strong memory with which it could match. The fillers also would not generate a strong memory signal because those fillers would also not strongly match Evan's weak memory.

Because the memory signals of different faces in the lineup are likely to be correlated, discriminability can be enhanced (therefore raising the empirical ROC) by using this correlation to facilitate the classification of a face as innocent or guilty. With simultaneous lineups, eyewitnesses seem to be able to take advantage of these correlated memory signals by automatically using an "ensemble" decision rule (Wixted et al., 2018). According to this idea, eyewitnesses first home in on the face that generates the strongest memory signal. The innocent-guilty classification decision and confidence are then based on how much that particular face stands out from the crowd (i.e., the difference between the face with the maximum memory signal and the average of all the memory signals).

Holding all else constant, increasing underlying discriminability (*d*′), will lead to a higher empirical ROC curve. The formula for d' is as follows:

$$d' = \frac{\mu_G - \mu_I}{\sigma} \tag{1}$$

μ_G is the mean memory strength of guilty suspects, μ_I is the mean memory strength of innocent suspects, and σ is the standard deviation for guilty and innocent suspects. This formula is the same formula used for calculating the Cohen’s *d* effect size for a 2-sample equal variance *t*-test. Performing the ensemble subtraction will reduce σ when the suspect and filler memory signals are correlated, which would therefore increase d'. Lineup procedures that allow eyewitnesses to better take advantage of correlated memory signals in their classification decision should therefore enhance discriminability. For this reason, if the confidence procedure encourages witnesses to consider each face in isolation, rather than utilizing correlated memory signals, it might lower the ROC relative to a traditional simultaneous lineup.

Yet, even if the confidence procedure does harm witnesses’ own abilities to discriminate innocent and guilty suspects, there is another intriguing possibility: it may still be possible to correct for this impairment by leaving the decision variable in the hands of the legal system. Valuable information from collected correlated memory signals (e.g., confidence ratings to filler faces) could be used computationally to improve discriminability in the confidence procedure. Increasing discriminability by reducing error variance operates in a similar manner to how covariates can be used to increase the effect size between two conditions of a research study (see Meyvis & Van Osselaer, 2018 for a discussion of increasing effect sizes by the use of covariates). Imagine that researchers are interested in whether a new type of mathematics instruction is an improvement over an old type of instruction. Participants are randomly assigned to receive either the new or old type of instruction. According to the logic of random assignment, the two groups should be equated *on average*. However, there is still likely to be considerable variability in the baseline mathematics skills between participants. One way that this between-participants variance can be reduced is by administering a pre-test to all participants before they complete the experimental manipulation. The dependent variable of interest could then be the difference in performance between pre- and post-test scores in the two conditions (similar to how the ensemble decision works in the mind of the eyewitness), or the post-test scores could be the dependent variable of interest with the pre-test scores included as covariates. Similarly, when we used a logistic regression to classify target-present and target-absent lineups based on the suspect confidence ratings in the Brewer et al. (2020) data, AUC was numerically higher when the average filler rating was included as a covariate in the model compared to when it was not (.663 versus .651). Thus, the ability of the procedure to discriminate between guilty and innocent suspects might be improved by using confidence ratings to filler faces to reduce error variance.

As new research is conducted using the confidence procedure, it will be interesting to see how the valuable information from correlated memory signals can best be utilized to get on the highest ROC curve. Will there ultimately be a tension between the types of lineup procedures that best allow eyewitnesses to take advantage of correlated memory signals (e.g., traditional simultaneous lineups) and the types of lineup procedures that aim to give the legal system a raw readout of those signals (e.g., the confidence procedure)? Perhaps, but the goal should be to find lineup procedures that yield the highest ROC curves. Whether this is accomplished by leaving the innocent-guilty classification decision in the hands of the eyewitness or the legal system is still unknown.

2. Develop a model to make predictions about scenarios not yet tested

Brewer et al. collected data from 1,669 adults and 273 children who viewed four different mock-crimes and subsequently made confidence judgments sequentially to 12 lineup members (2020). Brewer and Doyle (2021) summarize the findings in lay terms (see Table 1); For example, explaining that the probability of guilt was around .8 or .9 when the suspect received a maximum confidence value over 75%. A key question that remains is whether the pattern of findings holds over the range of system and estimator variables encountered in practice. For example, Brewer and Doyle (2021) query how the similarity of the fillers to the suspect might change the distribution of confidence judgments and evidence of guilt.

A model-based interpretation of findings would help to predict if the same conclusions would be expected to generalize to other scenarios (e.g., different *d'*, different placement of the decision criteria, etc.). A model-based approach would specify—mathematically—the theoretical basis for confidence judgments in the confidence procedure. As an example, the ensemble model assumes that witnesses identify the lineup face with the highest memory strength, if the difference between that lineup face and the average of all the lineup faces exceeds the witness's criteria (Wixted et al., 2018). Simulations using the ensemble model can be used to generate hypothetical witness identification outcomes in different scenarios not yet tested (e.g., when the lineup fillers are more or less similar-looking to the suspect; see Colloff et al., 2021). In the same way, a model for the confidence procedure—that specifies how the confidence judgement to each face is determined—could be used to simulate the pattern of confidence ratings to lineup members in scenarios not yet tested. Those simulations could also elucidate if the patterns in evidence of guilt observed for the four proposed rules by Brewer and colleagues (i.e., suspect unique max, suspect multiple max, suspect < unique max, suspect < multiple max; 2020, 2021) generalize across different conditions and which rules are most useful for determining suspect guilt.

Brewer and colleagues state that the confidence procedure is based on the theoretical premise that the confidence judgements are principally determined by the memory strength elicited by each face (e.g., Sauer et al., 2008). If this is the case, then faces that are perceived as more similar to the culprit's face stored in memory will receive a higher confidence rating, and each rating should be relatively immune to other contextual factors, such as the sequence of the other faces presented in the lineup, or their relative similarity (e.g., Sauer et al., 2008; Brewer et al., 2020).

However, other research suggests that it might be unlikely that lineup confidence judgements are immune to other contextual factors. For example, adding fillers that look nothing like the culprit increases witness confidence in choosing a person who looks like the culprit (Charman et al., 2011; Hanczakowski et al., 2014; Horry & Brewer, 2016; Windschitl & Chambers, 2004). One might intuit that the effect of contextual information is minimized in a sequential lineup, where each match-to-memory judgement is made one at a time, in the absence of other lineup members (Brewer & Doyle, 2021). However, there appear to be complex dependencies in sequential lineups (e.g., Gronlund et al., 2012; Horry et al., 2012; Rotello & Chen, 2016; Wilson et al., 2019). For example, Wilson et al. (2019) collected confidence ratings to each face in a 6-person sequential lineup using a scale ranging from −100 (sure the face was not seen) to +100 (sure this is the face of the culprit). Wilson et al. reported that participants were more certain that a subsequent filler was not the culprit (i.e., they were more conservative) after making an identification (rating the face closer to -100) than after making no identification (rating the face further from -100). This illustrates that later confidence ratings in sequential lineups may be influenced by earlier confidence judgements.

Given evidence that confidence judgements, even in sequential lineups, may not be as simple as a direct access to memory strength associated with each face, a more detailed understanding of the theoretical basis of confidence judgments in the confidence procedure would be beneficial. A model-based interpretation of the confidence judgements made during the confidence procedure would enable predictions to be made about how the evidence of guilt findings laid out by Brewer and Doyle (2021), may (or may not) generalize to other contexts.

Conclusion

We have outlined two key areas of work for further developing the confidence procedure as an alternative to the traditional identification procedure. Namely, we recommend that (1) research compare the confidence procedure to the traditional lineup procedure that it would replace, measuring the ability to discriminate innocent from guilty suspects, and (2) a model-based interpretation of confidence judgments be developed to help make predictions about evidence of guilt from the confidence procedure in different scenarios not yet tested. Once steps 1 and 2 have been achieved, more detailed consideration about implementation would be a later important step. A strong headwind faced by the confidence procedure is how to implement the procedure in police departments and aid interpretation for legal-decision makers. For example, asking a juror to believe that the suspect is likely guilty, even when the witness gave the highest rating to a filler, might be tricky. Generating discussion between academics and practitioners—as Brewer and Doyle have done in this issue—appears to be a productive way to move forward in this regard.

References

Brewer, N., & Doyle, J. (2021). Changing the face of police lineups: Delivering more information from witnesses. *Journal of Applied Memory in Research and Cognition.*

Brewer, N., & Wells, G. L. (2006). The confidence-accuracy relationship in eyewitness identification: Effects of lineup instructions, foil similarity and target-absent base rates. *Journal of Experimental Psychology: Applied, 12*(1), 11–30. https://doi.org/10.1037/1076-898X.12.1.11

Brewer, N., Weber, N., & Guerin, N. (2020). Police line-ups of the future? *American Psychologist, 75*(1), 76–91. https://doi.org/10.1037/amp0000465

Brewer, N., Weber, N., Wootton, D., & Lindsay, D. S. (2012). Identifying the bad guy in a lineup using confidence judgments under deadline pressure. *Psychological Science, 23*(10)*,*1208–1214. https://doi:10.1177/0956797612441217

Charman, S. D., Wells, G. L., & Joy, S. W. (2011). The dud effect: Adding highly dissimilar fillers increases confidence in lineup identifications. *Law and Human Behavior, 25,* 479–500. https://doi.org/10.1007/s10979-010-9261-1

Colloff, M. F., Flowe, H. D., Smith, H. J., Seale-Carlisle, T. M., Meissner, C. A., Rockey, J. C., Pande, B., Kujur, P., Parveen, N., Chandel, P., Singh, M. M., Pradhan, & S., Parganiha, A. (in press). Active exploration of faces in police lineups increases discrimination accuracy. *American Psychologist.*

Colloff, M. F., Wilson, B. M., Seale-Carlisle, T. M., & Wixted, J. T. (2021). Optimizing the selection of fillers in police lineups. *Proceedings of the National Academy of Sciences*, *118*(8) 8, e2017292118. https://doi.org/10.1073/pnas.2017292118

Gepshtein, S., Wang, Y., He, F., Diep, D., & Albright, T. D. (2020). A perceptual scaling approach to eyewitness identification. *Nature communications*, *11*(1), 1-10.

Gronlund, S. D., Carlson, C. A., Neuschatz, J. S., Goodsell, C. A., Wetmore, S. A., Wooten, A., & Graham, M. (2012). Showups versus lineups: An evaluation using ROC analysis. *Journal of Applied Research in Memory and Cognition, 1,* 221–228. https://doi.org/10.1016/j.jarmac.2012.09.003

Hanczakowski, M., Zawadzka, K., & Higham, P. (2014). The dud-alternative effect in memory for associations: Putting confidence into local context. *Psychonomic Bulletin & Review, 21,* 543–548.

Horry, R., & Brewer, N. (2016). How target–lure similarity shapes confidence judgments in multiple-alternative decision tasks. *Journal of Experimental Psychology: General, 145,* 1615–1634. https://doi.org/10.1037/xge0000227.

Horry, R., Palmer, M., & Brewer, N. (2012). Backloading in the sequential lineup prevents within-lineup criterion shifts that undermine eyewitness identification performance. *Journal of Experimental Psychology: Applied, 18,* 346–360. Doi: 10.1037/ a0029779.

Macmillan N. A. & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). Mahwah, NJ: Erlbaum.

Meyvis, T., & Van Osselaer, S. M. (2018). Increasing the power of your study by increasing the effect size. *Journal of Consumer Research, 44*(5), 1157–1173. https://doi.org/10.1093/jcr/ucx110

Mickes, L. (2015). Receiver operating characteristic analysis and confidence–accuracy characteristic analysis in investigations of system variables and estimator variables that affect eyewitness memory. *Journal of Applied Research in Memory and Cognition*, *4*(2), 93–102. https://doi.org/10.1016/j.jarmac.2015.01.003

Mickes, L., Flowe, H. D., & Wixted, J. T. (2012). Receiver operating characteristic analysis of eyewitness memory: comparing the diagnostic accuracy of simultaneous versus sequential lineups. *Journal of Experimental Psychology: Applied*, *18*(4), 361–376. https://doi.org/10.1037/a0030609

Mickes, L., Seale-Carlisle, T. M., Wetmore, S. A., Gronlund, S. D., Clark, S. E., Carlson, C. A., Goodsell, C. A., Weatherford, D., & Wixted, J. T. (2017). ROCs in Eyewitness Identification: Instructions versus Confidence Ratings. *Applied Cognitive Psychology*, *31*(5), 467–477. https://doi.org/10.1002/acp.3344

National Research Council (2014). *Identifying the Culprit: Assessing Eyewitness Identification.* Washington, DC: The National Academies Press.

Rotello, C. M., & Chen, T. (2016). ROC analyses of eyewitness identification decisions: An analysis of the recent debate. *Cognitive Research: Principles and Implications, 1,* 10. https://doi.org/10.1186/s41235-016-0006-7

Sauer, J. D., Brewer, N., & Weber, N. (2008). Multiple confidence estimates as indices of eyewitness memory. *Journal of Experimental Psychology: General, 137*(3), 528–547. https://doi.org/10.1037/a0012712

Sauer, J. D., Weber, N., & Brewer, N. (2012). Using ecphoric confidence ratings to discriminate seen from unseen faces: The effects of retention interval and distinctiveness. *Psychonomic Bulletin & Review, 19*(3), 490–498. https://doi.org/10.3758/s13423-012-0239-5

Seale-Carlisle, T. M., Wetmore, S. A., Flowe, H. D., & Mickes, L. (2019). Designing police lineups to maximize memory performance. Journal of Experimental Psychology: Applied, 25, 410–430. https://doi.org/10.1037/xap0000222

Smith, A. M., Yang, Y., & Wells, G. L. (2020). Distinguishing between investigator discriminability and eyewitness discriminability: A method for creating full receiver operating characteristic curves of lineup identification performance. *Perspectives on Psychological Science*, *15*(3), 589-607. https://doi.org/10.1177/1745691620902426

Starns, J., Cohen, A., & Rotello, C. (2021). Complete Method for Assessing the Effectiveness of Eyewitness Identification Procedures: Expected Information Gain. Retrieved from https://psyarxiv.com/az9xf/

Wilson, B. M., Donnelly, K., Christenfeld, N. J. S. & Wixted, J. T. (2019). Making Sense of Sequential Lineups: An Experimental and Theoretical Analysis of Position Effects. *Journal of Memory and Language, 104*, 108-125. https://doi.org/10.1016/j.jml.2018.10.002

Windschitl, P. D., & Chambers, J. R. (2004). The dud-alternative effect in likelihood judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30,* 198–215. https://doi.org/10.1037/0278-7393.30.1.198

Wixted, J. T. (2020). The forgotten history of signal detection theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 46*, 201-233. https://doi.org/10.1037/xlm0000732

Wixted, J. T., & Mickes, L. (2014). A signal-detection-based diagnostic-feature-detection model of eyewitness identification. *Psychological Review*, *121*(2), 262–276. https://doi.org/10.1037/a0035940

Wixted, J. T., Vul, E., Mickes, L. & Wilson, B. M. (2018). *Models of lineup memory. Cognitive Psychology, 105,* 81-114. https://doi.org/10.1016/j.cogpsych.2018.06.001