Journal of Applied Research in Memory and Cognition 10 (2021) 335–340

Contents lists available at ScienceDirect

Journal of Applied Research in Memory and Cognition

journal homepage: www.elsevier.com/locate/jarmac

Commentary

Use and Misuse of Receiver Operating Characteristic Analysis in Eyewitness Identification

Yueran Yang *

Department of Psychology, University of Nevada, USA

Sarah A. Moody

University of Nevada, USA

Keywords: Eyewitness identification, Receiver Operating Characteristic (ROC) analysis, Investigator operating characteristics, Expected utility

> "What is the aim of your policy? Without specifying an objective there is no way to evaluate any decision procedure."
> — Green and Swets (1966, p. 9)

Eyewitness evidence is often decisive in the police's judgment of a suspect's guilt (Wells et al., 2020). However, mistakes occur as eyewitness evidence is imperfect (e.g., Rakoff & Loftus, 2018). Innocent suspects can be judged as guilty, and guilty suspects can be judged as innocent. Such mistakes lead to serious consequences such as wrongful convictions of the innocent and the escape of the guilty. These serious consequences urge the legal system to understand and improve eyewitness practices and urge scholars to assist the legal system empirically.

Scholars have made significant efforts to improve eyewitness practices over the past few decades (Wells et al., 2020). However, a fundamental question remains unresolved: How should the legal system and scholars evaluate identification performance? Just as Green and Swets (1966) remarked in the opening quote, one cannot evaluate any policy procedures without a proper metric, and one cannot choose the proper metric without a well-specified objective. Therefore, the question of how to evaluate identification performance boils down to clarifying the legal system's objective.

So, what is the legal system's objective when considering different eyewitness practices? To increase eyewitnesses' discriminability? To change eyewitnesses' response bias? As Kovera and Evelo (2021) pointed out, eyewitness discriminability and response bias have arguably become the preeminent measures used in eyewitness research. They note their concern over the ubiquity of these measures, particularly at the expense of experimental methods that include examination of the social context in which identifications are situated. We concur with Kovera and Evelo's argument in favor of reintroducing socially situated paradigms into eyewitness research to improve identification practices. In addition, we contend the receiver operating characteristic (ROC) analysis should not be completely disregarded but differently used to achieve the legal system's objective. Kovera and Evelo recognize eyewitness discriminability and response bias can influence the extent to which the legal system can achieve its objective while also recognizing these measures do not directly reflect the legal system's objective.

We argue the misemphasis on eyewitnesses' operating characteristics arises from the misconception that the legal system is evaluating "eyewitness performance" (Wixted & Mickes, 2018). The legal system is certainly interested in eyewitness performance, but only to the extent that it influences investigator performance (Wells et al., 2015). Indeed, the legal system's objective concerns developing eyewitness practices that can improve *investigator performance*, as police investigators are the legal actors who collect and use eyewitness evidence (Smith et al., 2020). In ROC language, investigators are the *operators* who judge suspects' guilt using the information provided by eyewitness responses.

ROC Analysis in Eyewitness Identification

* Correspondence concerning this article should be addressed to Yueran Yang, Department of Psychology, University of Nevada, Reno, NV, United States. Contact: yuerany@unr.edu (Y.Y.).

Police investigators' use of eyewitness evidence can be deemed a classification problem. With eyewitness evidence, investigators are attempting to classify suspects as either guilty or innocent. Investigators can operate conservatively, classifying suspects as guilty only when witnesses identify suspects with high confidence. Alternatively, investigators can operate less conservatively, classifying suspects as guilty even when witnesses identify suspects with low confidence. Investigators can be even less conservative, classifying suspects as guilty even when witnesses reject a lineup.

Investigator Operating Characteristics

An ROC curve summarizes investigator operating characteristics, which are a collection of investigators' false and true positive rates when investigators use different criteria to make classification decisions. These pairs of false and true positive rates are called "operating points" on an ROC curve (Fawcett, 2006; Powers, 2012). Figure 1 Panel (a) shows a typical ROC curve in eyewitness research. Such an ROC curve is often referred to as a "partial" ROC curve because it only includes suspect identification rates at different confidence levels (Lampinen, 2016).

The "partial" ROC curve in Figure 1 Panel (a) illustrates investigators' false and true positive rates at three operating points. The ROC point labeled "IDS high" depicts investigators' false and true positive rates if investigators classify suspects as guilty when witnesses identify suspects with high confidence (and classify suspects as innocent when witnesses make other responses, including identifying suspects with medium or low confidence, identifying fillers, or rejecting lineups). The point "IDS medium" depicts investigators' false and true positive rates if investigators classify suspects as guilty as long as witnesses identify suspects with medium confidence (or high confidence). Similarly, the point "IDS low" depicts investigators' false and true positive rates if investigators classify suspects as guilty as long as witnesses identify suspects with low confidence (or medium or high confidence).

The two ending points, (0, 0) and (1, 1), are also valid operating points. The point (0, 0) reflects the situation when investigators *always* classify suspects as innocent, and the point (1, 1) reflects the situation when investigators *always* classify suspects as guilty, regardless of eyewitness evidence.

The assumption underlying the "partial" ROC curve is that investigators can only operate at these three points. In other words, it assumes investigators only use suspect identifications with different confidence levels as possible criteria to classify suspects' guilt. Practically, this assumption may be reasonable. Theoretically, it is limited.

Investigators can have more possible operating points as a lineup can generate more outcomes beyond suspect identifications (Smith & Ayala, 2021). For example, investigators can classify suspects as guilty even when witnesses identify fillers or reject lineups. If considering all possible operating points, the ROC curve will become what Smith et al. (2020) called the "full" ROC curve. Figure 1 Panel (b) shows such an ROC curve when adding filler identifications and rejections as possible operating points. The newly added point "IDF" depicts investigators' false and true positive rates if investigators classify suspects as guilty as long as witnesses identify fillers (or identify suspects). The point "REJ" depicts if investigators classify suspects as guilty as long as witnesses reject lineups (or identify fillers or suspects), that is if investigators *always* classify suspects as guilty. Of course, this ROC curve can have more operating points if investigators further differentiate filler identifications and rejections with witnesses' confidence levels (see Smith et al., 2020 for an example).

When comparing lineup practices, different ROC curves reflect investigator operating characteristics under these different practices. For example, Figure 2 displays the ROC curves using data from Colloff et al. (2016) to compare lineups with different filler similarities. For the sake of simplicity, we focused on two of the four conditions: the "do-nothing" condition in which the suspect had a distinctive facial feature but the fillers did not, and the "block" condition in which the suspect's

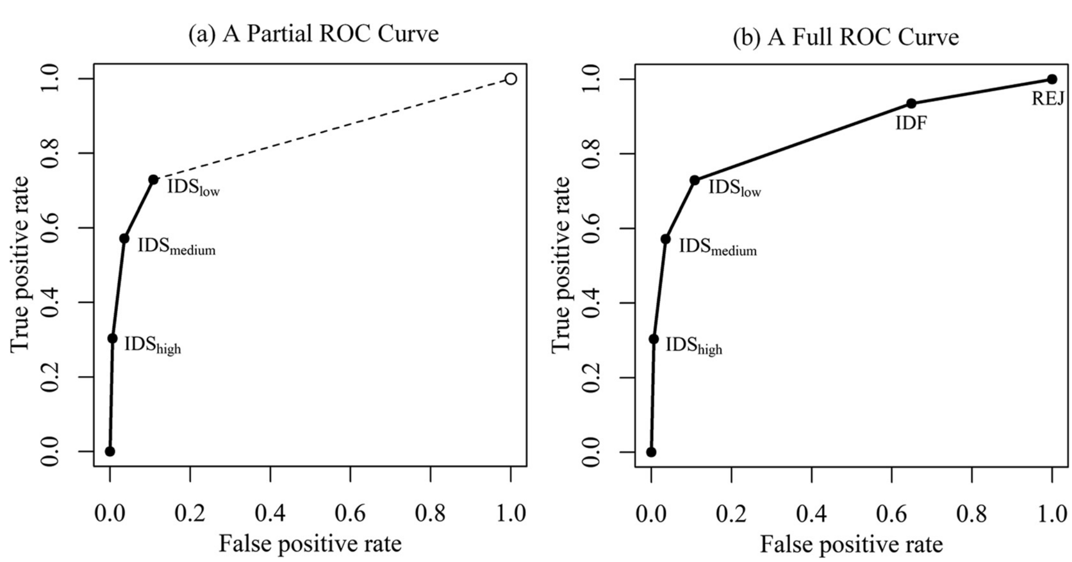

Figure 1. Partial and full ROC curves in eyewitness research.

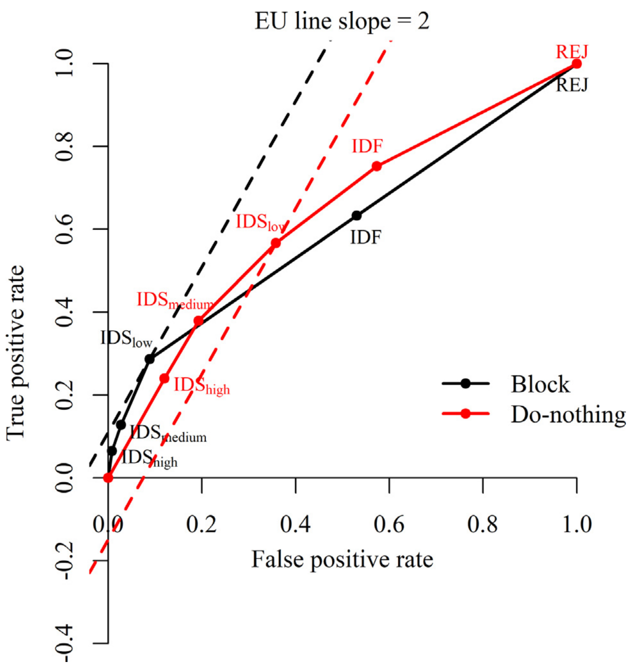

Figure 2. ROC curves for block and do-nothing lineups. The dashed lines show the expected utility lines (EU lines) assuming a utility-prior ratio of 2 (i.e., a slope of 2). The y-intercepts of the EU lines reflect the expected utilities of investigator classifications when investigators operate at the "IDS low" points using the "block" and "do-nothing" lineups, respectively. See the "Analyzing ROC Curves" section for more detail.

distinctive feature and the fillers' corresponding facial areas were blocked. We also collapsed some confidence levels (see Table 1).

The ROC curves in Figure 2 show the investigator operating characteristics under the "block" and "do-nothing" conditions. The assumption here is investigators can possibly use five operating criteria when judging suspects' guilt, which include suspect identifications with high, medium, and low confidence, filler identifications, and rejections. Readers can add more operating points onto the ROC curves if assuming investigators will further differentiate the eyewitness evidence. Readers can also remove some operating points if assuming investigators will never use them. For example, readers will get the "partial" ROC curves if assuming investigators will not judge suspects as guilty based on filler identifications or rejections (i.e., removing the "IDF" and "REJ" points in Figure 2).

Analyzing ROC Curves

As discussed above, an ROC curve reflects investigator operating characteristics but not eyewitness operating characteristics. But how should the legal system and scholars use ROC curves to select better lineup practices? A misconception is to compare the areas under the ROC curves (AUC) or the partial areas under the curves (pAUC; Gronlund et al., 2012; Mickes et al., 2012). For a continuous ROC curve, the AUC is often used as an index of "accuracy" because it equals the average true positive rate across the entire range of the false positive rate (Swets, 1988; Swets et al., 2000). Yet, this measure might not be proper for evaluating investigator performance due to several reasons. First, for a single ROC curve, false positive rates might not occur equally often at all possible operating points. Second, investigator ROC curves are discrete, meaning the false positive rates may not happen at the same levels between different practices (Lampinen, 2016; Smith et al., 2018). And third, using AUCs may over-simplify the comparisons, especially when ROC curves intersect (Adams & Hand, 1999).

A more proper measure is expected utility (Clark, 2012; Lampinen et al., 2019; Smith et al., 2019; Yang et al., 2019). The expected utility of investigator classification decisions is the weighted average of the utilities of all possible classification outcomes. The larger the expected utility, the better the performance.

In fact, ROC analysis relies on the expected utility metric (Yang, 2021). One can visualize the expected utility of investigator classification as a linear line in the ROC space. The dashed linear lines in Figure 2 show examples of such expected utility lines (EU lines). The slope of an EU line is determined by the prior probability of guilt and the utilities of investigator classification outcomes. The y-intercept of an EU line is a linear function of the expected utility when investigators operate at the ROC point through which the EU line goes (for more detail, see Yang, 2021).

For example, the EU lines in Figure 2 assume a slope of 2. This slope comes from an arbitrary choice that the ratio between the utilities of investigator classification outcomes is 10 (i.e., the Blackstone ratio; Volokh, 1997) and that the prior odds of guilt is 5 (i.e., the prior probability of guilt is 0.833; $prior\ probability\ of\ guilt = \frac{prior\ odds\ of\ guilt}{1+prior\ odds\ of\ guilt} = \frac{5}{1+5} = 0.833$). The slope is the ratio between these two quantities (i.e., the "utility-prior ratio;" see Yang, 2021). The black dashed line shows the EU line for the "block" lineups, and the red dashed

Table 1
True and False Positive Rates in Block and Do-Nothing Conditions from Colloff et al. (2016)

Evidence	Block			Do-nothing		
	True positive	False positive	DR	True positive	False positive	DR
IDS high (90–100)	0.065	0.008	8.18	0.240	0.120	2.00
IDS medium (70–80)	0.063	0.019	3.23	0.140	0.073	1.92
IDS low (0–60)	0.159	0.061	2.60	0.187	0.165	1.13
IDF (0–100)	0.346	0.442	0.78	0.186	0.215	0.86
REJ (0–100)	0.367	0.470	0.78	0.248	0.427	0.58

Note. DR = diagnosticity ratio.

line shows the EU line for the "do-nothing" lineups. Both EU lines go through the ROC point "IDS low" on their respective ROC curves, assuming investigators operate at the "IDS low" point no matter which lineup procedure they use. In other words, investigators will judge a suspect as guilty as long as the suspect is identified with low confidence (or medium or high confidence).

The EU lines convey three important messages. First, the intercept of the "block" EU line is larger than that of the "do-nothing" EU line. Because the expected utility is a linear function of the intercept, it indicates that the expected utility of the "block" lineups is larger than that of the "do-nothing" lineups if investigators operate at the "IDS low" points and if the utility-prior ratio is 2.

Second, the intercept of the "block" EU line has reached its maximum at the "IDS low" point among all possible operating points on the "block" ROC curve. But the intercept of the "do-nothing" EU line has not. The intercept of the "do-nothing" EU line can increase if investigators operate at a different point. Visually speaking, the EU line can move up to the "IDS high" point on the "do-nothing" ROC curve to maximize its intercept. In terms of expected utility, the "block" lineups have reached their maximized expected utility, but the "do-nothing" lineups can increase their expected utility if investigators operate at the "IDS high" point, in other words, if investigators judge a suspect as guilty only when the suspect is identified with high confidence.

Third, even if investigators operate at the "IDS high" point on the "do-nothing" ROC curve, the intercept of the maximized EU line (not shown in Figure 2) is still smaller than that of the "block" EU line. In other words, the maximized expected utility of the "do-nothing" lineups will still be smaller than that of the "block" lineups. Therefore, if operating at its optimal ROC point (i.e., the "IDS low" point), the "block" lineups will yield better outcomes than the "do-nothing" lineups no matter which ROC point the "do-nothing" lineups operate.

From the above example, an ROC curve can be useful for at least two purposes: to examine a lineup's *actual* expected utility and to locate a lineup's *maximized* expected utility. If investigators' operating points are fixed, one can map the EU lines onto the ROC space and then examine the intercepts. If investigators' operating points are not fixed, one can locate the optimal operating point for investigators to maximize expected utility. The optimal operating point is the one that yields the largest EU line intercept among all possible EU lines.

In this sense, an ROC curve reflects a lineup's *potential* to maximize the expected utility of investigator classification decisions. However, whether a lineup can achieve its potential depends on whether investigators are operating at the optimal point. Even if a lineup can potentially maximize the expected utility to a larger extent than other lineups, it is not necessarily the case that the lineup's *actual* expected utility will be larger. For example, if operating at the "IDF" point, the "block" lineups will yield a smaller expected utility than the "do-nothing" lineups. After all, the *actual* expected utility also depends on the ROC point at which investigators operate.

It is important to note the above conclusions are based on the assumption that the utility-prior ratio is 2 (i.e., the EU lines have a slope of 2). The conclusions may change if the utility-prior ratio changes. However, the analytical procedures stay the same. One can either examine the actual expected utilities or the maximized expected utilities for any lineup practices. If investigators' operating points are fixed, one can locate the EU lines according to the utility-prior ratio and investigators' operating points. If investigators' operating points are not fixed, one can locate the EU lines that can maximize the expected utility and then examine the intercepts. Larger intercepts entail larger expected utilities.

In short, ROC analysis evaluates investigator performance in terms of expected utility. Once the utility-prior ratio is estimated, EU lines can visualize the expected utilities when investigators operate at different ROC points for any lineup practice. The connection between an ROC curve and expected utility answers our opening question—what is the legal system's objective when considering different eyewitness practices? The objective is to maximize the expected utility when investigators use eyewitness evidence to make classification decisions. We argue this is the objective underlying ROC analysis.

Factors Influencing Investigator Performance

ROC curves and EU lines jointly provide a useful tool for analyzing and comparing investigator performance when investigators use eyewitness evidence to judge suspects' guilt. Such analysis relies on the expected utility metric. As the objective is to increase expected utility, it would be useful to examine ways to achieve this objective beyond ROC analysis.

Figure 3 summarizes factors that influence the expected utility of investigator classification decisions. As shown in the dashed rectangle, the current eyewitness research focuses primarily on the "evidence generation stage," that is, how different variables influence eyewitness responses, thereby influencing investigator operating characteristics (Fulero, 2009; Wells, 1978). As Kovera and Evelo (2021) pointed out, two underlying psychological mechanisms are of particular interest: eyewitness memory and social context. These two mechanisms are shown as the dashed textboxes and arrows in Figure 3, meaning they may serve as mediators that can possibly explain the effects of different variables on investigator operating characteristics. Understanding these mechanisms, therefore, is critical for developing better lineup procedures to improve investigator performance.

Nevertheless, developing better lineup procedures is not the only way to improve investigator performance. ROC analysis suggests investigator performance relies not only on investigator operating characteristics but also on investigators' actual operating points (Swets et al., 2000; Yang, 2021). A lineup procedure may have great potential to maximize expected utility but will not achieve its potential unless investigators choose a proper criterion for making classification decisions. Therefore, it is also important for the legal system and scholars to consider the "evidence usage stage," which concerns how investigators should use eyewitness evidence.

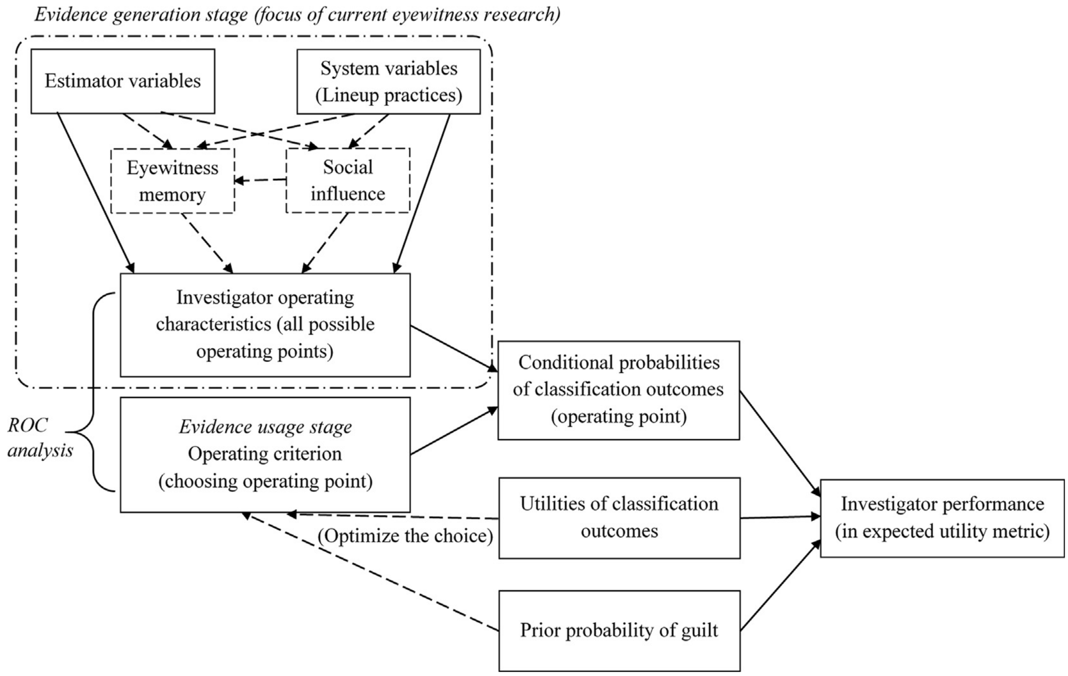

Figure 3. Factors influencing investigator performance.

As shown in Figure 3, using ROC analysis still does not capture the whole picture. ROC analysis focuses primarily on changing the conditional probabilities of classification outcomes to increase expected utility. However, expected utility also depends on the prior probabilities and the utilities of classification outcomes. In particular, the legal system can possibly change the prior probability of guilt by requiring different amounts of incriminating evidence before conducting a lineup (Wells et al., 2020; Yang et al., 2019), thereby changing the expected utility. As such, the legal system and scholars must consider these parameters as well.

In this sense, developing better lineup procedures is far from sufficient for the legal system to achieve its objective—to maximize the expected utility of investigator classification decisions. After all, expected utility depends not only on investigator operating characteristics but also on many other parameters, such as the investigator's actual operating point, the prior probability of guilt, and the utilities of investigator classification outcomes. The legal system needs to carefully analyze and, if possible, control these parameters. If the legal system just focuses on developing better lineup procedures without carefully considering other parameters involved in evaluating investigator performance, the long-term outcomes may not be desirable.

Conclusion

Kovera and Evelo (2021) noted the importance of understanding social context effects on eyewitness behavior as well as social context's decline in eyewitness research due to the rise of ROC analysis. We echo their concerns. ROC analysis should not dim the importance of social factors in eyewitness research. Instead, the proper use of ROC analysis should consider how social context influences eyewitness behavior, which in turn influences the legal system's objective to maximize the expected utility of investigator classification decisions when using eyewitness evidence.

We argue that two misconceptions have hindered the proper use of ROC analysis in eyewitness research. The first misconception is that an ROC curve describes eyewitness operating characteristics. Rather, an ROC curve describes investigator operating characteristics when investigators use eyewitness evidence to judge suspects' guilt (Smith et al., 2020). This explanation applies to both partial and full ROC curves, which differ only in their assumptions on investigators' operating points.

The second misconception is that ROC analysis relies on AUCs to assess performance. Rather, EU lines should be used jointly with ROC curves to quantify and compare investigator performance (Yang, 2021). EU lines can map the expected utility of investigator classification decisions onto an ROC space for any utility-prior ratio and any operating point. The relation between an ROC curve and expected utility reveals the objective underlying ROC analysis—to maximize the expected utility of investigator classification decisions.

ROC analysis's reliance on the expected utility metric also enlightens other possibilities to improve investigator performance beyond just developing alternative lineup procedures (i.e., changing investigator operating characteristics). The legal system also needs to consider other parameters involved in estimating the expected utility, including investigators' actual operating points, the prior probability of guilt, and the utilities of classification outcomes.

Author Statement

We appreciate the support of the National Science Foundation, SES 2017046, for this research project.

Author Contributions

Y.Y. developed the research idea. Both Y.Y. and S.M. contributed to the writing of the manuscript. Both authors approved the final version of the manuscript for submission.

Conflict of Interest

The authors declare that they have no conflict of interest.

References

Adams, N. M., & Hand, D. J. (1999). Comparing classifiers when the misallocation costs are uncertain. *Pattern Recognition, 32*(7), 1139–1147. https://doi.org/10.1016/S0031-3203(98)00154-X.

Clark, S. E. (2012). Costs and benefits of eyewitness identification reform: Psychological science and public policy. *Perspectives on Psychological Science, 7*(3), 238–259. https://doi.org/10.1177/1745691612439584.

Colloff, M. F., Wade, K. A., & Strange, D. (2016). Unfair lineups make witnesses more likely to confuse innocent and guilty suspects. *Psychological Science, 27*(9), 1227–1239, https://doi.org/10.1177%2F0956797616655789.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010.

Fulero, S. M. (2009). System and estimator variables in eyewitness identification: A review. In D. A. Krauss & J. D. Lieberman (Eds.), *Psychological expertise in court* (pp. 57–78). Ashgate.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. Wiley.

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*(4), 221–228. https://doi.org/10.1016/j.jarmac.2012.09.003.

Kovera, M. B., & Evelo, A. J. (2021). Eyewitness identification in its social context. *Journal of Applied Research in Memory and Cognition, 10*(3), 313–327.

Lampinen, J. M. (2016). ROC analyses in eyewitness identification research. *Journal of Applied Research in Memory and Cognition, 5* (1), 21–33. https://doi.org/10.1016/j.jarmac.2015.08.006.

Lampinen, J. M., Smith, A. M., & Wells, G. L. (2019). Four utilities in eyewitness identification practice: Dissociations between receiver operating characteristic (ROC) analysis and expected utility analysis. *Law and Human Behavior, 43*(1), 26–44. https://doi.org/10.1037/lhb0000309.

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.

Powers, D. M. (2012). ROC-ConCert: ROC-based measurement of consistency and certainty. *2012 spring congress on engineering and technology*, (pp. 1–4) IEEE. https://doi.org/10.1109/SCET.2012.6342144.

Rakoff, J. S., & Loftus, E. F. (2018). The intractability of inaccurate eyewitness identification. *Daedalus, 147*(4), 90–98. https://doi.org/10.1162/daed_a_00522.

Smith, A. M., & Ayala, N. T. (2021). Do traditional lineups undermine the capacity for eyewitness memory to rule out innocent suspects?. *Journal of Applied Research in Memory and Cognition, 10*(2), 215–220. https://doi.org/10.1016/j.jarmac.2021.03.003.

Smith, A. M., Lampinen, J. M., Wells, G. L., Smalarz, L., & Mackovichova, S. (2019). Deviation from perfect performance measures the diagnostic utility of eyewitness lineups but partial area under the ROC curve does not. *Journal of Applied Research in Memory and Cognition, 8*(1), 50–59. https://doi.org/10.1016/j.jarmac.2018.09.003.

Smith, A. M., Wells, G. L., Smalarz, L. A., & Lampinen, J. M. (2018). Increasing the similarity of lineup fillers to the suspect improves the applied value of lineups without improving memory performance: Commentary on Colloff, Wade, and Strange (2016). *Psychological Science, 29*(9), 1548–1551. https://doi.org/10.1177/0956797617698528.

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 in Psychological Science, 15*(3), 589–607. https://doi.org/10.1177/1745691620902426.

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science, 240*(4857), 1285–1293. https://doi.org/10.1126/science.3287615.

Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest, 1*(1), 1–26, https://doi.org/10.1111%2F1529-1006.001.

Volokh, A. (1997). N guilty men. *University of Pennsylvania Law Review, 146*, 173–216.

Wells, G. L. (1978). Applied eyewitness testimony research: System variables and estimator variables. *Journal of Personality and Social Psychology, 36*(12), 1546–1557. https://doi.org/10.1037/0022-3514.36.12.1546.

Wells, G. L., Kovera, M. B., Douglass, A. B., Brewer, N., Meissner, C. A., & Wixted, J. T. (2020). Policy and procedure recommendations for the collection and preservation of eyewitness identification evidence. *Law and Human Behavior, 44*(1), 3–36. https://doi.org/10.1037/lhb0000359.

Wells, G. L., Yang, Y., & Smalarz, L. (2015). Eyewitness identification: Bayesian information gain, base-rate effect-equivalency curves, and reasonable suspicion. *Law and Human Behavior, 39* (2), 99–122. https://doi.org/10.1037/lhb0000125.

Wixted, J. T., & Mickes, L. (2018). Theoretical vs. empirical discriminability: The application of ROC methods to eyewitness identification. *Cognitive Research: Principles and Implications 3*, (1), 9. https://doi.org/10.1186/s41235-018-0093-8.

Yang, Y. (2021). *Evaluating classification performance: Receiver operating characteristic and expected utility,* [Unpublished manuscript]. Department of Psychology, University of Nevada.

Yang, Y., Smalarz, L., Moody, S. A., Cabell, J. J., & Copp, C. J. (2019). An expected cost model of eyewitness identification. *Law and Human Behavior, 43*(3), 205–219. https://doi.org/10.1037/lhb0000331.

Received June 21, 2021
Received in revised form June 23, 2021
Accepted June 24, 2021