Journal of Applied Research in Memory and Cognition 4 (2015) 324–328

Contents lists available at ScienceDirect

Journal of Applied Research in Memory and Cognition

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

Reply

ROC analysis of lineups obscures information that is critical for both theoretical understanding and applied purposes

Gary L. Wells [a,*], Andrew M. Smith [b], Laura Smalarz [c]

[a] Iowa State University, Ames, IA, United States
[b] Queen's University, Kingston, ON, Canada
[c] Williams College, Williamstown, MA, United States

ARTICLE INFO

Article history:
Received 10 August 2015
Accepted 30 August 2015
Available online 16 September 2015

Keywords:
Eyewitness identification
Lineups
Showups
Receiver operating characteristic (ROC) analysis
Bayesian analysis
Lineup fillers
Lineup discriminability
Filler siphoning
Diagnosticity

ABSTRACT

Our previous article (Wells et al., 2015a. *Journal of Applied Research in Memory and Cognition*) showed how ROC analysis of lineups does not measure underlying discriminability or control for response bias. Wixted and Mickes (2015. *Journal of Applied Research in Memory and Cognition*) concede these points. Hence, in this article we focus more on how forcing the 3×2 lineup into the 2×2 structure required for ROC analysis obscures important underlying phenomena of theoretical value. Moreover, ROC analysis fails to account for the unique diagnostic properties of exonerating eyewitness behaviors (filler identifications and rejections). We describe how an examination of the full 3×2 structure helps reveal the critical underlying phenomena that ROC analysis hides. We also show how a Bayesian approach yields a family of diagnosticity functions that exposes the unique diagnosticity of all three eyewitness behaviors (suspect identifications, filler identifications, and rejections). Moreover, we show how Bayesian methods can examine diagnosticity as a function of witness confidence for all three eyewitness behaviors, which gives it a significant applied advantage over ROC analysis.

We discussed the problems with the use of Receiver operating characteristic (ROC) curves for analyzing lineup data in our previous article (Wells, Smalarz, & Smith, 2015a, hereafter called "our previous article"). Because ROC proponents force the natural 3×2 structure of lineups into a 2×2 representation by treating false positive identifications of fillers as if they were rejections, ROC analysis cannot possibly measure underlying discriminability from lineups[1]. Wixted and Mickes (2015) now state explicitly that ROC analyses of lineups do not measure underlying discriminability. In fact, Wixted and Mickes say that the idea that ROC analysis is to examine underlying discriminability independently of response bias "needs to be nipped in the bud" (p. 15).

Although Wixted and Mickes' (2015) latest article explicitly acknowledges that ROC analyses on lineups do not assess underlying discriminability, their body of work has implied otherwise. Wixted and Mickes (2014) wrote "The lineup procedure that yields the highest empirical ROC is the one that theoretically best facilitates the discrimination between guilty and innocent suspects *by reducing the overlap between the corresponding memory strength distributions*" (p. 402, emphasis added). More recently, Mickes and Wixted (2015) explained that "discriminability refers to the ability to distinguish the face that was seen in the video from faces that were not" (p. 402). Moreover, the introductions to their prior articles are steeped in Signal Detection Theory and the concept of separating signal from noise and not once do they acknowledge that their ROC curves do not measure underlying discriminability. Wixted and Mickes (2014) even launched a theory about underlying discriminability on the basis of their ROC results on lineups (diagnostic-feature detection theory, or DFD theory), whereby the presence of lineup fillers (rather than showing a face in isolation) enhances discriminability because it permits witnesses to "attach weight to features that might be diagnostic while discounting features that are nondiagnostic" (p. 269).

If Wixted and Mickes (2015) did not mean that ROC analyses on lineups were to be interpreted as measures of underlying discriminability, then that is certainly the impression they have given the field. Indeed, following on the lead of Wixted and Mickes (2014), Wetmore et al. (2015) clearly interpreted their ROC results on lineups versus showups as evidence of better underlying discriminability for lineups (see our first article). And the National

* Corresponding author. Tel.: +1 515 294 6033.
E-mail address: glwells@iastate.edu (G.L. Wells).

[1] For problems that naturally fit the 2×2 structure, on the other hand, we believe that ROC analysis is an appropriate measure of underlying discriminability.

http://dx.doi.org/10.1016/j.jarmac.2015.08.010

Research Council (NRC) (2014), relying on Wixted and Mickes' writings regarding ROC analyses of lineups and Wixted's presentation to their eyewitness committee, said "the ROC approach possesses a distinct advantage because the dimensions of analysis – discriminability and response bias – map directly onto the mechanistic parameters of causal models of human memory. . .. In other words, the [ROC] approach affords insight in and quantification of the sensory and cognitive processes that are believed to underlie memory-based classification decisions, such as eyewitness identifications" (p. 87). Yes, the National Research Council (NRC) report got it wrong by interpreting ROC analyses on lineups as measures of underlying discriminability. But that is how the NRC eyewitness committee read and interpreted Wixted and Mickes' work.

Although Wixted and Mickes (2015) have now distanced themselves from this view, the assumption that ROC analysis managed to get at underlying discriminability was the primary luster and allure that drew people into the ROC fold and was the basis on which ROC proponents argued that ROC analysis was superior to other approaches.

1. The continued dismissal of the importance of filler identifications

Despite acknowledging that ROC analysis on lineups does not measure discriminability, Wixted and Mickes (2015) continue to dismiss the importance of filler identifications. As we described in our previous article, ROC analysis forces the 3×2 structure of lineups into a 2×2 structure by collapsing false positive filler identifications into the category of rejections. In a strange attempt to dismiss the importance of this misclassification, Wixted and Mickes produced ROC curves with a dataset for which it would not have mattered whether filler identifications were counted as rejections or as false positives in terms of which procedure produced the higher ROC curve. But just because something holds in a selected example does not mean that it is generally true. Here, we show a striking counter to their claim.

We used data on biased versus fair lineups using the same method that Wixted and Mickes (2014) advocate (filler identifications treated as rejections) versus a method in which filler identifications are counted as false-positive responses in culprit-absent lineups (which is known as the Location ROC, or LROC, method). We used the pROC statistical package (Robin et al., 2011) to create ROC and LROC curves from Wetmore et al.'s (2015) data on fair versus biased lineups. As is clear in Fig. 1, the fair lineup has the higher ROC curve for the standard ROC (top Panel) but the biased lineup has the higher curve for the LROC (bottom Panel); a total reversal of which curve is the highest.

Neither the ROC curves nor the LROC curves in Fig. 1 reflect underlying discriminability or control for response bias. Our point here is simply to counter the misleading suggestion in Wixted and Mickes' (2015) example (their Fig. 3) that treating fillers as rejections versus false positives does not change conclusions about which of two lineup procedures has the higher ROC curve. Fig. 1 shows that how fillers are treated can actually reverse which procedure has the higher area under the curve.

2. Only a proper 3 × 2 examination can reveal the filler-siphoning process

Our previous article showed that the actual underlying process that explains how fair lineups reduce mistaken identifications of innocent suspects relative to showups and biased lineups is neither improved discriminability nor a change in response criterion. Instead, the process is filler siphoning. Specifically, good lineup fillers manage to siphon false-positive responses away from an

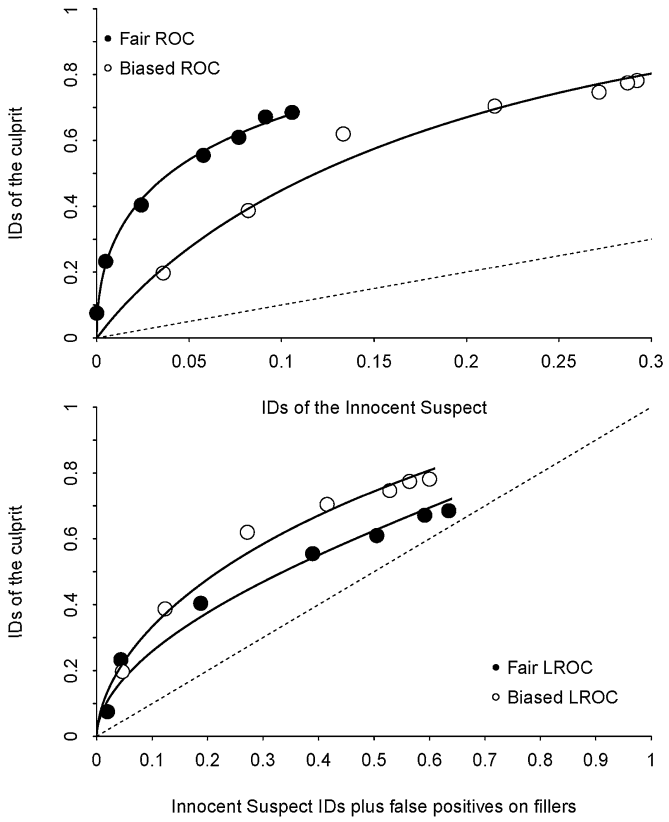

Fig. 1. ROC curves for fair versus biased lineups. *Note*: Data are from Wetmore et al.'s (2015) fair and biased lineups, moderate-similarity innocent suspect, collapsed over immediate and delay conditions.

innocent suspect in a culprit-absent lineup. Filler siphoning cannot occur at all with showups (because there are no fillers) and filler siphoning is suppressed with biased lineups (because the fillers do not resemble the culprit as well as the innocent suspect does). Our previous article also described why good fillers tend to not siphon away identifications of the culprit in a culprit-present lineup.

The central role of filler siphoning is apparent, however, only if one examines the entire 3×2 array and it is totally hidden by the 2×2 representation that ROC proponents force on the data. This filler-siphoning process is extremely important and is fundamental to our understanding of lineup data. Hence, even if ROC proponents have abandoned the idea that ROC analysis assesses underlying discriminability, the ROC approach masks our ability to see and understand one of the most fundamental processes that is operating with lineups.

3. Diagnosticity ratios assess all three eyewitness behaviors

Our previous article noted how ROC analysis examines only two of the six cells in the 3×2 lineup structure (correct and false suspect identification rates). The diagnosticity-ratio approach (derived from Bayes' Theorem), however, has a long history of examining all six cells because each of the three possible witness behaviors (suspect identifications, filler identifications, and rejections) has its own diagnosticity ratio (Wells & Lindsay, 1980). The diagnosticity of an identification of the suspect is the ratio of the probability of an accurate identification of the culprit (from culprit-present conditions) to the probability of a mistaken identification of the innocent suspect (from culprit-absent conditions). The diagnosticity of a rejection is the ratio of the probability of rejection when the suspect is not the culprit to the probability of rejection when the suspect is the culprit. The diagnosticity of a filler identification is the ratio of the

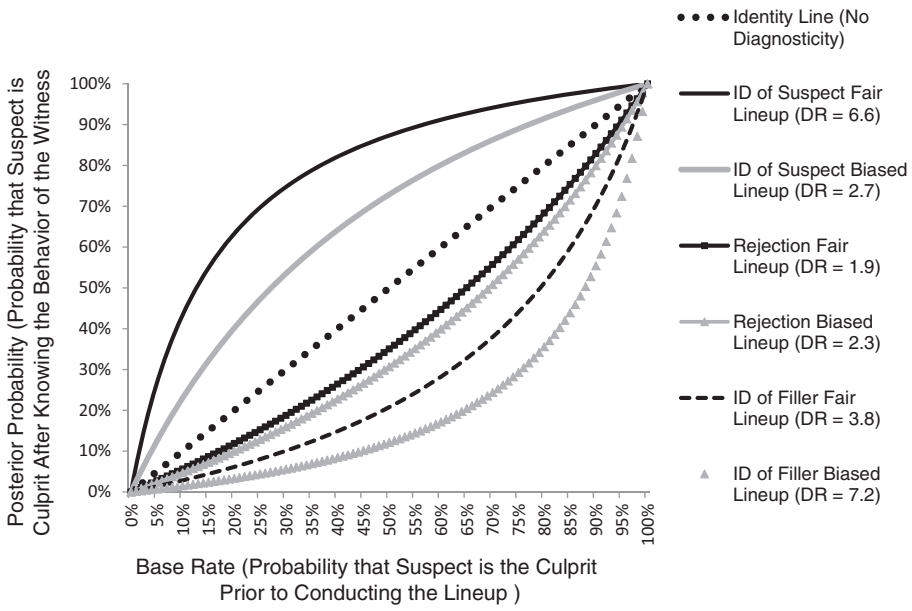

Fig. 2. Bayesian prior-by-posterior curves for all three witness behaviors for fair versus biased lineups. *Note*: Data are from Wetmore et al.'s (2015) fair and biased lineups, moderate-similarity innocent suspect, collapsed over immediate and delay conditions.

probability of a filler identification when the suspect is not the culprit to the probability of a filler identification when the suspect is the culprit (Wells & Lindsay, 1980; Wells & Turtle, 1986; Wells, Yang, & Smalarz, 2015b).

Despite a 35-year history of the diagnosticity ratio approach examining all three eyewitness behaviors, Wixted and Mickes (2015) nevertheless said "ROC analysis ignores filler IDs to the same extent that a conventional analysis on the diagnosticity ratio does. Both approaches have been based on correct and false ID rates computed from suspect IDs, and both have ignored filler IDs to the same extent" (p. 7). In other words, they are claiming that the diagnosticity/Bayesian approach only focuses on two of the six possible cells just as ROC analysis does. This is simply false.

In order to illustrate how the Bayesian diagnosticity approach utilizes all six cells, we used the Wetmore et al. (2015) data on fair versus biased lineups to create what are known in Bayesian terms as prior-by-posterior curves, as shown in Fig. 2. Each of the three witness behaviors (suspect IDs, filler IDs, rejections) has a unique curve (reflecting the diagnosticity ratio) relating the probability that the suspect is the culprit *before* conducting the lineup (prior probability) on the *X*-axis to the probability that the suspect is the culprit *after* knowing the behavior of the witness (posterior probability) on the *Y*-axis.

The dotted line in Fig. 2, called the identity line, is where these curves would fall if the behaviors were non-diagnostic of guilt (a diagnosticity ratio of 1.0). The distance of a curve from this identity line represents how much the behavior of the witness changes the probability that the suspect is the culprit. Curves above the identity line are incriminating (increasing the probability that the suspect is the culprit) and curves below the identity line are exonerating (decreasing the probability that the suspect is the culprit). Not surprisingly, an identification of the suspect increases the probability that the suspect is the culprit and rejections decrease the probability that the suspect is the culprit. Notice, however, that the fair lineup does a better job at incriminating the suspect (given an identification of the suspect) than does the biased lineup, whereas the biased lineup does a better job at exonerating the suspect (given a rejection) than does the fair lineup. We noted in our previous article that one procedure can be more diagnostic for identifications of the suspect while the other procedure is more diagnostic for rejections. ROC analyses cannot observe such a phenomenon.

Particularly interesting is the fact that filler identifications decrease the probability that the suspect is the culprit. In other words, the identification of a filler is an exonerating piece of information, a finding that has been consistently observed since it was first noted 35 years ago using Bayesian diagnosticity curves (Wells & Lindsay, 1980). The interpretation of this consistent finding is rather straightforward: When a witness identifies a filler, the witness is (in effect) saying "this filler looks more like the culprit than does the suspect" which, of course, is more likely when the suspect is innocent than when the suspect is the culprit. In fact, as can be seen in Fig. 2, the identification of a filler from a biased lineup was more diagnostic of innocence than the identification of a suspect from a fair lineup was diagnostic of guilt. None of this unique diagnostic value of filler identifications can be discovered by ROC analysis.

4. Diagnosticity ratios generate a family of confidence curves

ROC proponents continue to tout lineup ROC analyses because they track the data across levels of confidence. But, the Bayesian diagnosticity approach does (and has done) this as well (e.g., Wells et al., 2015b). We illustrate this by using the lineup versus showup data from Wetmore et al. (2015). We split the data into three levels of the 7-point confidence scale used by Wetmore et al. (low = 1–3, moderate = 4–5, high = 6–7) and plotted prior-by-posterior curves for each confidence level. The top panel of Fig. 3 shows the resulting family of diagnosticity curves for the lineup, which follow a neat order from more diagnostic to less diagnostic as one moves from higher to lower levels of confidence. This same pattern is largely repeated for rejections and for filler identifications (i.e., both tend to provide stronger evidence of innocence when they are made with higher levels of confidence). The bottom panel of Fig. 3 shows the results for showups. Notice that the curves are much closer to the diagonal identity line for showups than for lineups, which of course, indicates that showups are relatively less diagnostic than are lineups. The point of Fig. 3 is to illustrate that the Bayesian diagnosticity ratio can readily display diagnosticity as a function of confidence.

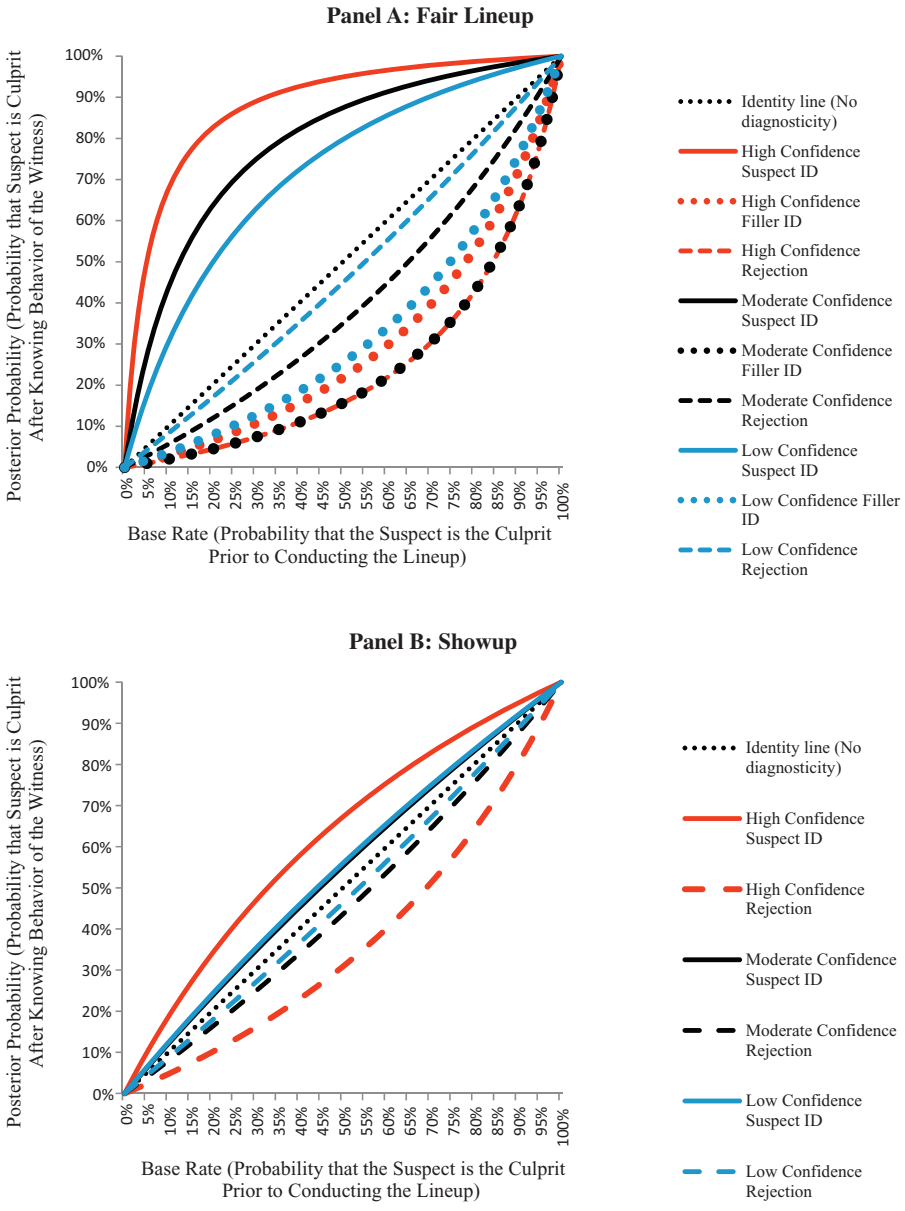

Fig. 3. Bayesian prior-by-posterior curves for a fair lineup and for a showup. *Note*: Data are from Wetmore et al.'s (2015) fair and biased lineups, moderate-similarity innocent suspect, collapsed over immediate and delay condition.

5. The diagnosticity approach would not lead to increasingly-conservative procedures

One argument against the Bayesian diagnosticity ratio for identifications of the suspect is that the ratio gets predictively larger when responding is more conservative. And the argument then is that the diagnosticity ratio approach would lead down a slippery slope to advocating increasingly-conservative responding until almost no witness makes an identification. But, this misrepresents the diagnosticity/Bayesian approach. The diagnosticity/Bayesian approach does not rely solely on diagnosticity for *identifications of the suspect*, which increase with increased conservatism. Unlike the ROC approach, the diagnosticity/Bayesian approach looks at *all three* witness responses to a lineup. That is important because with more conservative responding the diagnosticity ratio for exonerating behaviors goes *down*. In other words, increasing conservatism makes witnesses' rejection decisions *less* probative of innocence.

Because the diagnosticity approach looks at all six cells of the 3×2 lineup structure, this broader focus (that looks separately at the diagnosticity of rejections) provides resistance against any slippery slope toward advocating increasingly-conservative procedures. Like most slippery-slope arguments, which are easy to make, there are counter-veiling factors that tend to get ignored in making the argument.

6. Policy makers are interested in the diagnosticity ratio

Wixted and Mickes (2015) dismiss the relevance of the diagnosticity ratio by claiming "it is not the question of interest to policymakers" (p. 12). We are not sure who Wixted and Mickes think policy makers are or what kinds of things they think policy makers consider relevant. Eyewitness policy task-forces and commissions are comprised of judges, prosecutors, defense attorneys, and law enforcement, all of whom have an interest in diagnosticity: Judges make judicial policy about the rules of evidence and admissibility on the basis of the probative value of evidence (which is what the diagnosticity ratio reflects); Prosecutors have an

interest in law enforcement using procedures that permit them to trust identifications that are proffered as evidence in criminal cases (i.e., the diagnosticity ratio); Police detectives are interested in adopting procedures that give them good ratios of accurate to mistaken identifications (indicated by the diagnosticity ratio).

One of the most important policy-relevant contributions of the Bayesian approach is the discovery that filler identifications are diagnostic of the innocence of the culprit (Clark & Wells, 2008; Wells & Lindsay, 1980; Wells & Olson, 2002; Wells & Turtle, 1986; Wells et al., 2015b). This is profoundly relevant to policy because over 1/3rd of law enforcement agencies in the U.S. admit that they do not even write a report of the lineup if the witness identifies a filler (Police Executive Research Forum, 2013). Hence, the Bayesian approach has provided critical evidence for implementing policies that require clear records of the outcome of a lineup (and assessments of witness confidence) regardless of the witness behavior (suspect identification, filler identification, or rejection). Note that the critical data to support this policy cannot be observed with ROC analyses.

Mickes (2015) argued that a Bayesian/diagnosticity approach is appropriate for estimator variables but not for system variables. There are many reasons to disagree with that statement, not the least of which is how it misrepresents the system/estimator variable distinction. Specifically, the system/estimator relation is asymmetric because an estimator variable is not necessarily a system variable, but every system variable is an estimator variable (see Wilford & Wells, 2013 for a discussion of this). Lineup bias, for example, is a system variable but it is also an estimator variable in the sense that lineup bias is just as relevant to estimating the likely accuracy of an eyewitness as is an estimator variable such as cross-race identifications.

7. Understanding of lineups requires methods that use a 3 × 2 representation

We have shown some advantages for the diagnosticity/Bayesian approach to lineups. But, the more general point is that further advances in our understanding of the psychological processes operating in lineups requires that we examine the entire 3×2 matrix. Forcing the 3×2 structure of lineups into a 2×2 representation hides important underlying processes. We noted in our first paper, for example, how the 2×2 ROC approach led Wetmore et al. (2015) to posit that the lineup was better than the showup because the lineup engages an underlying psychological process as described in Wixted and Mickes' (2014) DFD theory, which posits a process of underlying discriminability. But when examining the full 3×2 array it is apparent that the lineup versus showup data actually go *against* DFD theory and that the actual explanation is something altogether different (i.e., filler siphoning).

We are still at a rudimentary level of understanding the underlying psychological processes involved in lineups and only through a better understanding of these psychological processes can we devise even better ways to improve the outcomes of lineups. Minimally that better understanding requires that we examine all six cells of the 3×2 structure of lineups rather than forcing the 3×2 into a 2×2 that confounds false positive identifications of fillers with rejections.

Conflict of interest statement

The authors declare that they have no conflict of interest.

Acknowledgements

Aspects of this work were supported by grant SES0850401 from the National Science Foundation to the lead author.

References

Clark, S. E., & Wells, G. L. (2008). On the diagnosticity of multiple-witness identifications. *Law and Human Behavior*, *32*, 406–422. http://dx.doi.org/10.1007/s10979-007-9115-7

Mickes, L., & Wixted, J. T. (2015). On the applied implications of the "verbal overshadowing effect". *Psychological Science*, *10*, 400–403. http://dx.doi.org/10.1177/1745691615576762

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*, 93–102. http://dx.doi.org/10.1016/j.jarmac.2015.01.003

National Research Council. (2014). *Identifying the culprit: Assessing eyewitness identification*. Washington, DC: The National Academies Press.

Police Executive Research Forum. (2013). *A national survey of eyewitness identification procedures in law enforcement agencies*. Police Executive Research Forum. Retrieved from ⟨http://policeforum.org/library/eyewitness-identification/NIJEyewitnessReport.pdf⟩.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J. C., et al. (2011). pROC: An open-source package for R and S+ to analyze and compute ROC curves. *BMC Bioinformatics*, *12*, 77. http://dx.doi.org/10.1186/1471-2105-12-77

Wells, G. L., & Lindsay, R. C. L. (1980). On estimating the diagnosticity of eyewitness nonidentifications. *Psychological Bulletin*, *88*, 776–784. http://dx.doi.org/10.1037/0033-2909.88.3.776

Wells, G. L., & Olson, E. (2002). Eyewitness identification: Information gain from incriminating and exonerating behaviors. *Journal of Experimental Psychology: Applied*, *8*, 155–167. http://dx.doi.org/10.1037//1076-898X.8.3.155

Wells, G. L., & Turtle, J. W. (1986). Eyewitness identification: The importance of lineup models. *Psychological Bulletin*, *99*, 320–329, http://dx.doi.org/10.1037/0033-2909.99.3.320.

Wells, G. L., Smalarz, L., & Smith, A. M. (2015). ROC analysis of lineups does not measure underlying discriminability and has limited value. *Journal of Applied Research in Memory and Cognition*, *4*, 313–317.

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*, 99–122. http://dx.doi.org/10.1037/lhb0000125

Wetmore, S. A., Neuschatz, J. S., Gronlund, S. D., Wooten, A., Goodsell, C. A., & Carlson, C. A. (2015). Effect of retention interval on showup and lineup performance. *Journal of Applied Research in Memory and Cognition*, *4*, 1–7. http://dx.doi.org/10.1016/j.jarmac.2014.07.003

Wilford, M. M., & Wells, G. L. (2013). Eyewitness system variables. In B. L. Cutler (Ed.), *Reform of eyewitness identification procedures* (pp. 23–43). Washington, DC: American Psychological Association.

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

Wixted, J. T., & Mickes, L. (2015). Evaluating eyewitness identification procedures: ROC analysis and its misconceptions. *Journal of Applied Research in Memory and Cognition*, *4*, 318–323.