Eyewitness Memory

Laura Mickes[1] & John T. Wixted[2]

[1]University of Bristol, [2]University of California, San Diego

Laura Mickes, School of Psychological Science, University of Bristol, laura.mickes@bristol.ac.uk. John T. Wixted, Department of Psychology, University of California, San Diego, jwixted@ucsd.edu.

Abstract

In some respects, the story of eyewitness memory has remained unchanged for decades. For example, because memory is malleable, eyewitness memory evidence (like all types of forensic evidence) can be contaminated. Although the field of psychology has appreciated the malleability of memory for decades, the legal system's failure to appreciate that fact has led to many wrongful convictions. In other respects, the story of eyewitness memory – particularly eyewitness identification from a lineup – has changed in recent years. We focus on that story here. For decades, the field mistakenly believed that sequential lineups were diagnostically superior to simultaneous lineups, that confidence was, at best, only weakly predictive of accuracy, and that suboptimal estimator variables (e.g., short exposure, weapon focus, etc.) reduced the reliability of eyewitness identifications. But none of that is true, which raises a question: What went wrong? The chief problem seems to be that, long ago, applied psychology became divorced from basic experimental psychology, particularly with respect to recognition memory. As a result, eyewitness identification research did not make use of theories from basic science that have proven their ability to accurately guide the interpretation of data, particularly signal detection theory. In recent years, signal detection theory has been applied to a variety of eyewitness identification issues. As it turns out, simultaneous lineups are superior to sequential lineups, initial confidence (before contamination) is highly predictive of accuracy, and estimator variables are largely irrelevant once initial confidence is taken into account.

Eyewitness Memory

Intuition is usually a faithful servant, but it can sometimes lead an entire field astray. Consider, for example, how easy it is to convince someone (including you, probably) that sequential lineups are diagnostically superior to simultaneous lineups. In years gone by, lineups consisted of several individuals paraded in front of an eyewitness who would observe them through a one-way mirror. These days, instead of live lineups, police in the U.S. routinely use *photo lineups* (Police Executive Research Forum, 2013). A proper photo lineup consists of one photo of the suspect (who is either innocent or guilty) and five or more photos of "fillers" who are known to be innocent but who physically resemble the perpetrator (Wells, Kovera, Douglass, Brewer, Meissner, & Wixted, 2020). The photos can be shown to the eyewitness all at once (simultaneous lineup) or one at a time (sequential lineup).

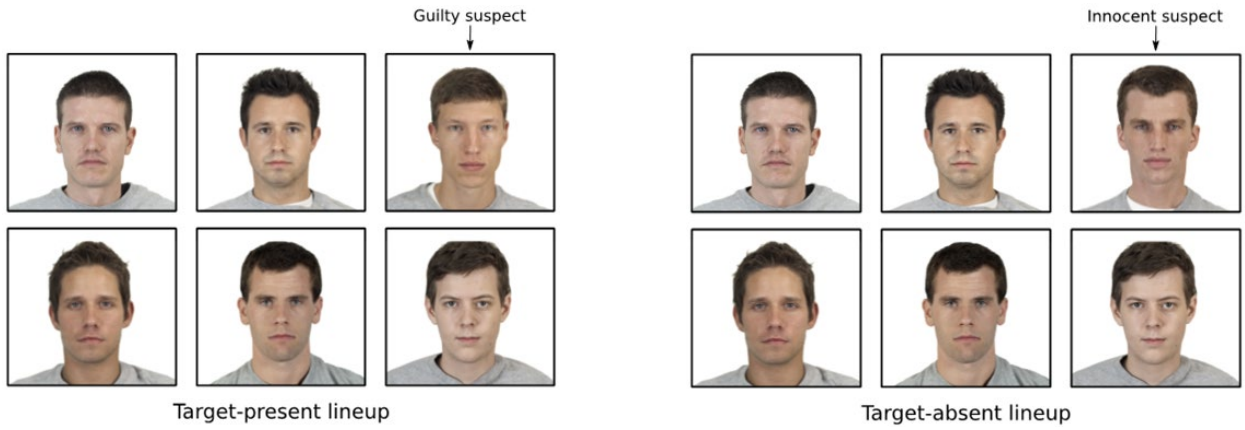

Figure 1. The left panel illustrates a target-present lineup containing a photo of a guilty suspect and five similar fillers. The right panel illustrates a target-absent lineup containing a photo of an innocent suspect and five similar fillers. In a real investigation, the police do not know if the suspect in the lineup is innocent or guilty (which is why the eyewitness is asked to identify the perpetrator if present), but in mock-crime studies, the experimenter does know and so can count the number of eyewitnesses who correctly choose the guilty suspect as well as the number who incorrectly choose the innocent suspect. Instead of choosing the suspect, eyewitnesses can pick a filler (which is always an error) or reject the lineup (a correct response for target-present lineups and an incorrect response for target-present lineups). Suspect faces and filler faces are from the Chicago Face database (Ma, Correll, & Wittenbrink, 2015).

In the seminal and most influential lab study comparing simultaneous and sequential lineups (cited 752 times according to Google Scholar in March 2020), Lindsay and Wells (1985) found that 58% of guilty suspects were correctly identified from target-present lineups (i.e.,

lineups containing the guilty suspect), and 43% of innocent suspects were misidentified from target-absent lineups (i.e., lineups containing an innocent suspect). This pattern of data—namely, the hit rate (HR) = .58 and false alarm rate (FAR) = .43—reflects rather poor performance, but the task was intentionally arranged to be difficult by using an innocent suspect who resembled the perpetrator more than the fillers did. When a sequential lineup was used instead of a simultaneous lineup, 50% of guilty suspects were correctly identified from target-present lineups, and only 17% of innocent suspects were misidentified from target-absent lineups. Most people—including you, probably—would agree that the small decrease in the HR (letting a few guilty perpetrators escape justice) is a small price to pay in order to achieve a large decrease in the FAR (protecting many innocent suspects from being wrongfully convicted).

If results like these convince you that sequential lineups are superior, then you are making an interpretive mistake, and it is here that a viable model of memory-based decision-making becomes essential. The great value of such a model is that it protects us from the dangerous intuitions that we will otherwise almost certainly have, and the model that has served that role more effectively than any other for over half a century is signal detection theory (Wixted, 2019). As described in more detail later, signal detection theory frees you from the mental trap you might find yourself in right now, namely, the trap of believing that a lineup procedure is adequately characterized by a single pair of hit and false alarm rates. In other words, it frees you from mistakenly thinking that the choice is between one lineup procedure that is constrained to yield HR/FAR values of .58/.43 vs. another that is constrained to yield .50/.17. In truth, the two procedures are each constrained to yield an *entire family* of HR and FAR values, and the analytical methodology used to measure each procedure's diagnostic accuracy must take that fact into account. Later, we rely on signal detection theory to consider how to do that, and

when we do, it will become clear that simultaneous lineups are diagnostically superior to sequential lineups (for theoretically sensible reasons).

The example presented above involves a choice that a policymaker, such as a police chief, has to make. Specifically, as a matter of policy, should a police department require its officers to administer photo lineups simultaneously or sequentially? A completely different question faces judges and jurors who must assess the reliability of a particular eyewitness who has identified a defendant on trial for having committed a crime (Mickes, 2015). This is not a question about policy. Instead, it is a question about the reliability of eyewitness memory. If you are like a lot of people, you believe that eyewitness memory is inherently unreliable. And why not? Seminal research by Elizabeth Loftus beginning in the 1970s found that false memories are easier to implant than anyone previously realized (e.g., Davis & Loftus, 2018; Loftus, Miller & Burns, 1978; Loftus & Palmer, 1974; Loftus & Pickrell, 1995; Loftus, 2003). Indeed, the malleability of memory has been so well established and so frequently written about that we do not address that issue in detail here (see adjacent *Sidebar* for a brief summary of the key issues). In addition, it is now well known that mistaken eyewitness identification played a key role in many wrongful convictions. Beginning in the early 1990s, the Innocence Project has used DNA testing to establish that many convicted prisoners are in fact innocent, and in approximately 71% of these cases, an eyewitness not only misidentified an innocent defendant in front of a judge and jury, they did so with high confidence (Garrett, 2011; Innocence Project, 2019). If any more evidence of the unreliability of eyewitness memory were needed, further confirmation was provided by research apparently showing that even in stress-free and procedurally ideal laboratory lineup tests, misidentifications are common and confidence is only weakly indicative of accuracy. Summarizing the relevant research, Penrod and Cutler (1995) explained that

SIDEBAR

Perhaps the most well-known illustration of the malleability of memory is the "misinformation effect." In a classic experiment reported by Loftus, Miller and Burns (1978), participants viewed a series of slides depicting successive stages of an auto-pedestrian accident. The critical slide showed a car stopped at either a stop sign or a yield sign. After viewing the slides, the participants answered a series of questions, one of which casually mentioned the sign depicted in the critical image. For some participants, the question was consistent with what they had seen (e.g., "Did another car pass the red Datsun while it was stopped at the stop sign?") whereas for other participants, the question was inconsistent with what they had seen (e.g., "Did another car pass the red Datsun while it was stopped at the yield sign?"). When given a subsequent recognition test involving a choice between two slides, one showing a car stopped at a stop sign and another showing a car stopped at a yield sign, the participants who received consistent information chose the correct slide 75% of the time. By contrast, the participants who received inconsistent information chose the *incorrect* slide 59% of the time. Thus, with surprising ease, people can be led to believe that they saw something they did not actually see. In a later classic study showing how elaborate a false memory can be, participants were induced to falsely remember they had become lost in a shopping mall as a child (Loftus & Pickrell, 1995).

Findings like these led to a deeper understanding of troubling legal dramas that played out in the 1980s and 1990s. For example, during the 1980s, a moral panic over day-care sexual abuse was later attributed to the unintentional implantation of false memories in young children during suggestive interviews (Ceci & Bruck, 1993; Ceci, Loftus, Leichtman, & Bruck, 1994). Similarly, during the repressed-memory epidemic in the 1990s, adult patients in psychotherapy recovered childhood memories of having been sexually abused by their relatives, some of whom were prosecuted based on that evidence alone. Only later did it become clear that the apparently recovered memories may have been unintentionally implanted by psychotherapists as they repeatedly probed a patient's childhood memories using techniques such as "guided imagery" (Loftus & Ketcham, 1994).

The once surprising realization that memory is malleable—so much so that extremely elaborate false memories can be unintentionally implanted—is now established knowledge. A related line of research further revealed that memory errors initially associated with low confidence can be transformed into high-confidence memory errors if feedback is provided suggesting that the memory was correct (e.g., Wells & Bradfield, 1998). All of these findings underscore the importance of testing memory using proper procedures and early in a police investigation, thereby minimizing the chances of distortion and/or contamination.

confidence "…is a weak indicator of eyewitness accuracy even when measured at the time an ID is made and under relatively 'pristine' laboratory conditions" (p. 830).

If errors made with high confidence are common even for initial IDs made under pristine testing conditions, then it would be fair to conclude that eyewitness memory is inherently unreliable. However, such a conclusion would also be completely at odds with signal detection theory. Thus, something must be wrong, either with the theory that has successfully guided thinking about recognition memory for more than half a century or with the analyses that led to the conclusion that eyewitness memory is unreliable even under good testing conditions. As we shall see, and in agreement with signal detection theory, eyewitness memory is highly reliable the first time it is tested in the sense that high confidence implies high accuracy and low confidence implies low accuracy. Moreover, because the act of testing memory contaminates memory, there is no second identification test that will help to determine the guilt or innocence of a defendant. All later tests, including the one that occurs in court in front of a jury (indeed, *especially* that test), should be regarded as contaminated forensic evidence.

Even if you are prepared to accept the possibility that, on an initial test, eyewitnesses are more reliable than once believed, you might nevertheless assume, as many do, that the witnessing conditions have to be ideal for that to be true. However, in practice, witnessing conditions are rarely ideal. For example, the witness might only get a brief look at the perpetrator (short exposure), or the distance between the witness and the perpetrator might be large (long distance), or the race of the perpetrator might differ from that the eyewitness (cross-race), or the attention of the witness might be drawn to a weapon (weapon focus), or the witness might be extremely anxious (high stress). On top of all of these potentially deleterious encoding factors, it is not uncommon for a long time to pass between the commission of the witnessed crime and the

administration of the photo lineup test (long retention interval). Worse, for many crimes, more than one of these variables – known as "estimator variables" – is unfavorable. Estimator variables are factors that affect the accuracy of eyewitness memory but that are not under the control of the legal system (Wells, 1978).

The estimator variables listed above are known to reduce the accuracy of eyewitness memory. That being the case, if you were a conscientious judge or juror attempting to assess the reliability of an eyewitness who identified the suspect with high confidence on the initial (uncontaminated) lineup test, you might want to know all about the estimator variables. After all, it seems reasonable to suppose that if (for example) the accuracy of a high-confidence ID made under perfect encoding conditions is 95% correct, then it must be considerably lower than that if the encoding conditions were less-than-perfect for one or more of the reasons listed above. Indeed, for decades, eyewitness experts have testified to that effect in courts of law, but, once again, now for the third time, intuitive conclusions like these simply do not follow once the problem is conceptualized in terms of signal detection theory. As it turns out, high-confidence IDs are less *likely to occur* given poorer encoding conditions (understandably), but those that do occur tend to remain highly reliable.

In what follows, we focus on how and why thinking in the field of eyewitness memory has changed in recent years (Gronlund & Benjamin, 2017). In no small part, this change was spurred by a report from the National Academy of Sciences on eyewitness identification (National Research Council, 2014), which helped to bring basic scientists with expertise in sensory and cognitive processes into the picture. As the co-chairs of that committee recently put it, these scientists provided what had previously been lacking: "…a principled mechanistic understanding of how people see, remember, and make decisions" (Albright & Rakoff, p. 21).

The result, in our view, is a story that differs dramatically from what was long understood to be true. Whereas the old message was to use sequential lineups, ignore confidence (even on the first test), and pay close attention to estimator variables, the new message is to use simultaneous lineups, pay extremely close attention to initial confidence, and ignore the estimator variables. The new message would not have been appreciated absent the guidance provided by signal detection theory, so we begin there.

Signal Detection Theory

The roots of signal detection theory date back to Fechner (1860), but it did not emerge in its modern form until 1953 (Green & Swets, 1966; Macmillan & Creelman, 2005; Wixted, 2019). In psychology, it was first applied to visual and auditory psychophysical tasks. The prototypical psychophysical detection task involves a mixture of stimulus-present trials (e.g., a dim light is flashed) and stimulus-absent trials (no light is flashed), and the observer's job is to indicate whether or not a stimulus was presented. A hit occurs when the observer correctly responds "Yes" on stimulus-present trials, and a false alarm occurs when the observer incorrectly responds "Yes" on stimulus-absent trials.

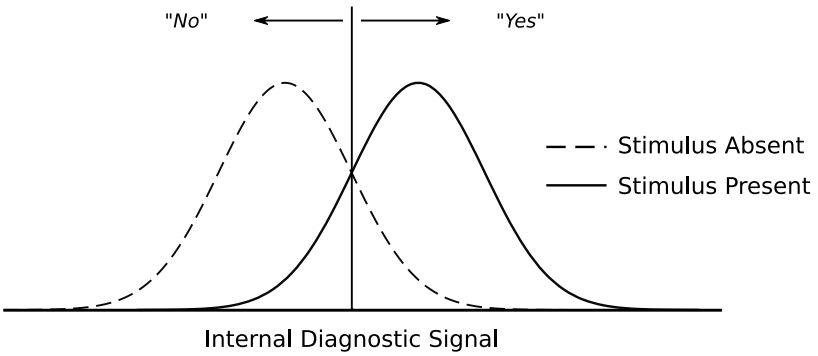

Figure 2. An illustration of the basic signal detection model for a simple yes/no task. Theoretically, the subject checks an internal diagnostic signal (e.g., sensations obtained by monitoring a sensory channel) in an effort to determine whether or not a stimulus was presented on a given trial. A noisy signal is returned, the average value of which is higher on signal (i.e., stimulus-present) trials than noise (i.e., stimulus-absent) trials. A decision criterion, the placement of which is under the control of the subject, is used to determine if the signal on a given trial is strong enough to support a "yes" decision (otherwise, the decision is "no"). For simplicity, the signals are usually assumed to follow equal-variance Gaussian distributions, but those assumptions are not inherent to signal detection theory (e.g., a different distributional form can be assumed).

As illustrated in Figure 2, although the *decision* is binary (yes or no), the internal diagnostic signal upon which the decision is based is continuous. In other words, sensations generated by neural activity in the monitored sensory channel are not all-or-none but instead range continuously from low to high across trials. Moreover, the sensation on a given trial arises from spontaneous neural activity (i.e., neural noise) plus, on stimulus-present trials, additional neural activity generated by the presentation of the stimulus. Thus, theoretically, the average sensation is stronger on stimulus-present trials compared to stimulus-absent trials, and the distribution of sensations across trials is assumed to be Gaussian in form. Unless the task is arranged to be trivially easy, the distributions overlap to some degree (as they do in Figure 2), which means that no level of sensation can be relied upon to perfectly discriminate stimulus-present from stimulus-absent trials. Thus, a decision criterion must be placed on the sensation axis such that any sensation greater than the criterion yields a "yes" decision; otherwise, the decision is "no."

In Figure 2, the criterion is placed directly between the two distributions, though it need not be placed there. Instead, some observers might place it more to the left (i.e., in a more liberal location, yielding higher hit and false alarm rates); others might place it more to the right (i.e., in a more conservative location, yielding lower hit and false alarm rates). No matter where the criterion is placed, the ability to discriminate the two states of the world is given by d', which is the distance between the means of the two distributions in terms of standard deviation units. The distributions shown in Figure 2 are two standard deviations apart, so $d' = 2$.

The signal detection framework was extended to recognition memory by Egan (1958). The prototypical recognition task involves the presentation of a list of items (e.g., words or faces) followed by a test in which the items are presented one at a time for a yes/no recognition

decision. Usually, half the test items are old (i.e., targets that appeared on the list) and half are new (i.e., lures that did not appear on the list). In other words, the *base rate* of targets and lures are typically equal, though they need not be. For this task, the *x*-axis in Figure 2 (labelled "Internal Diagnostic Signal") now represents a continuous memory signal (e.g., the degree to which a test item matches a representation in episodic memory) instead of a continuous perceptual signal, but otherwise it is the same model that applies to a psychophysical task.

The showup signal detection model. The basic list-memory task is conceptually analogous to a common eyewitness identification procedure known as a showup. In a showup, the eyewitness is presented with only one individual, who is either innocent (a lure) or guilty (the target). Thus, there is only one item on the "list" in this case (the perpetrator), and each eyewitness is tested only once (with a suspect, who is either innocent or guilty). The signal detection interpretation of showup performance is illustrated in Figure 3. It is essentially the

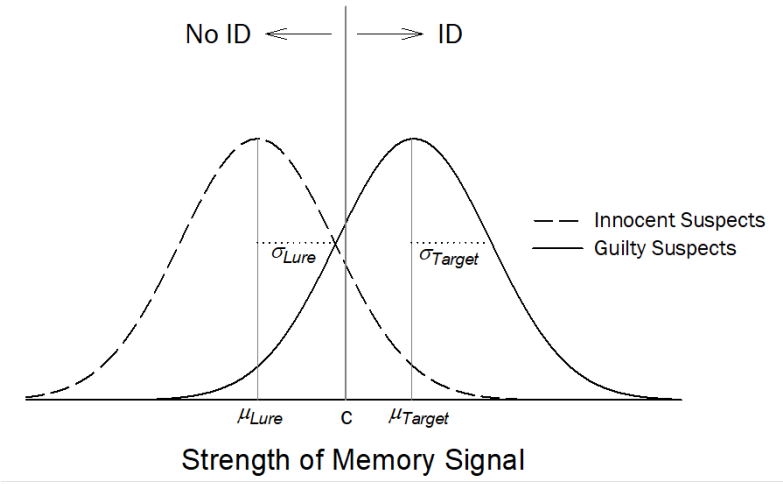

Figure 3. The standard signal detection model depicted in Figure 2 except now conceptualized as being applied to an eyewitness ID procedure known as a *showup*. In a showup, the eyewitness is tested with a suspect only (innocent or guilty), that is, without the fillers that are included when a lineup is used (Figure 1). The target is the guilty suspect (with the mean and standard deviation of the guilty suspect distribution being μ_{Target} and σ_{Target}, respectively), and lure is the innocent suspect (with the mean and standard deviation of the innocent suspect distribution being μ_{Lure} and σ_{Lure}, respectively).

same general signal detection model shown in Figure 2 except that the continuous diagnostic signal is now specifically a memory signal, and the variance of the target and lure distributions now includes new sources of variance, such as stimulus differences within each stimulus category (e.g., different innocent suspects will generate different memory signals) and individual differences over eyewitnesses (i.e., some witnesses will form stronger memories than others). As depicted here, the memory signals generated by targets (guilty suspects) and lures (innocent suspects) are distributed according to Gaussian distributions with means of μ_{Target} and μ_{Lure}, respectively, and standard deviations of σ_{Target} and σ_{Lure}, respectively (with $\sigma_{Target} = \sigma_{Lure}$ for simplicity). As before, the difference between the target and lure means in terms of their common standard deviation is the main signal-detection-based measure of discriminability, d', except that now it is a population measure of discriminability, not a measure of discriminability for any particular participant.

A great virtue of the signal detection framework is that it also allows one to easily conceptualize confidence in a recognition decision. More specifically, confidence in an ID corresponds to the highest confidence criterion exceeded by the memory strength associated with a given face, whether it is a target or a lure. The model shown in Figure 4 illustrates this point. In this illustration, the overall decision criterion is no longer placed directly between the two distributions. Instead, it is shifted to the left (to a more liberal setting), and additional confidence criteria have been added to the picture. As depicted here, the model corresponds to a 5-point confidence scale (1 = low confidence → 5 = high confidence), and each confidence rating is associated with its own decision criterion. For example, a face that generates a memory signal that falls to the extreme right of the horizontal memory-strength axis will not only be identified but will be identified with high confidence (5) because its strength falls above c_5. Later, we

describe how this conceptualization naturally predicts a strong relationship between confidence and accuracy, but for now, we are still developing the story of how best to compare the diagnostic accuracy of competing lineup procedures.

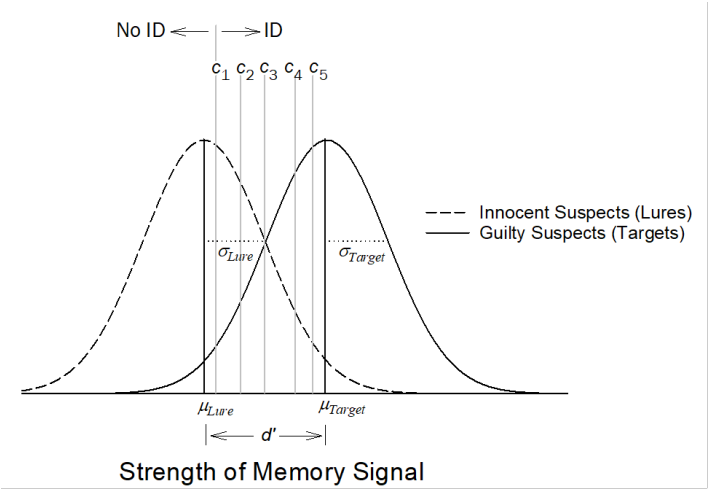

Figure 4. The same eyewitness ID signal detection model depicted in Figure 3, with additional decision criteria added to illustrate how the model interprets IDs made with different levels of confidence.

The basic signal detection model for lineups. The models shown in Figures 3 and 4 apply to a showup recognition test in which the eyewitness is presented with a single test face, either a target (guilty suspect) or a lure (innocent suspect). The same basic model applies to a lineup, but the way it works is somewhat different (Wixted & Mickes, 2014; Wixted, Vul, Mickes & Wilson, 2018). As noted earlier (Figure 1), a photo lineup consists of a picture of one suspect (the person who the police believe may have committed the crime) plus several additional photos of physically similar foils (i.e., fillers) who are known to be innocent. A target-present lineup includes the perpetrator along with (usually 5) similar appearing foils; a target-absent lineup is the same except that the perpetrator is replaced by an innocent suspect.

As illustrated in Figure 5, a 6-item target-present lineup is conceptualized as 5 random draws from what we now refer to as the Foil distribution and 1 random draw from the Target

distribution; a 6-item target-absent lineup is conceptualized as 6 random draws from the Foil distribution (all statistically independent of each other in the simplest case). Note that the memory strength of the innocent suspect (referred to earlier as the lure) is represented by one of the values drawn from the Foil distribution because, if the lineup is constructed in such a way that the innocent suspect does not stand out (i.e., if the lineup is fair), the innocent suspect is, from the point of view of the witness, just another individual who fits the description of the perpetrator but who did not actually commit the crime (i.e., the innocent suspect is just another foil).

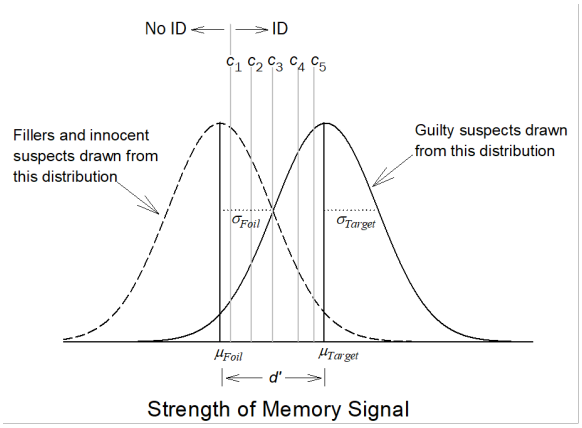

Figure 5. The standard eyewitness ID signal detection model illustrated in Figure 4 except now conceptualized as applying to a simultaneous lineup.

Because multiple faces are involved on a given lineup test, more than one face may exceed the decision criterion. How does the witness decide whether or not to make an ID? The simplest decision strategy would be for the witness to first determine the photo that generates the strongest (MAX) memory signal and to then identify that person (who is either the suspect or a filler) if the signal exceeds a decision criterion. If it exceeds the decision criterion, the MAX face would be identified even if other faces in the lineup also generate a memory signal that exceeds the decision criterion. The stronger the memory signal associated with the MAX face is, the

higher the confidence in that ID would be. If no face in the lineup generates a memory signal that exceeds the criterion, the lineup would be rejected (i.e., no ID would be made).

As with a showup, the HR corresponds to the proportion of the target-present lineups in which the guilty suspect is correctly identified (i.e., the proportion of target-present lineups in which the suspect yields the MAX signal and the strength of that signal exceeds the decision criterion), and the FAR corresponds to the proportion of target-absent lineups in which the innocent suspect is incorrectly identified (i.e., the proportion of target-absent lineups in which the suspect yields the MAX signal and the strength of that signal exceeds the decision criterion). To completely characterize lineup performance, one must also compute the Filler ID rate for both target-present and target-absent lineups, as well as the No-ID rate from both. The target-present and target-absent Filler ID rates are the proportion of lineups (target-present and target-absent, respectively) in which a filler is incorrectly identified, and the target-present and target-absent No-ID rates are the proportion of lineups (target-present and target-absent, respectively) that are incorrectly rejected. Thus, lineup performance is characterized by the Suspect ID rates, the Filler ID rates and the No-ID rates for target-present and, separately, target-absent lineups.

Although Filler ID rates and No-ID rates are important for model-fitting purposes (i.e., the appropriate signal detection model is the one that can accurately characterize Suspect ID rates, Filler ID rates and No-ID rates from both target-present and target-absent lineups), for applied purposes, the focus is on Suspect ID rates (i.e., the HR and FAR). The applied goal is to minimize the FAR while simultaneously maximizing the HR, regardless of what the Filler ID rates and No-ID rates might be when that state of affairs is achieved (Wixted & Mickes, 2018). Models are critical for figuring out how to achieve that goal, but the ultimate test is a purely empirical one.

Alternative signal detection models for lineups. It is important to note here that, because of the presence of fillers, there are several different basic signal detection models for lineups. The lineup model described above is technically known as the Independent Observations model (Wixted et al., 2018). Its simplicity makes it a useful model for conceptualizing lineup performance. However, a seemingly odd feature of this model is that if the memory signals associated MAX face and the next-best face are both very strong, confidence in the ID of the MAX face would nevertheless be high (because the decision is based on the strength of the MAX face considered in isolation). An alternative and, we argue, more viable signal detection model for lineups—known as the Ensemble model—holds that the memory-strength variable upon which the decision is based is not the raw memory signal associated with the MAX face. Instead, it is the degree to which the MAX memory signal stands out from the average memory signal generated by the faces in the lineup. If that difference score exceeds a decision criterion, then the MAX face is identified, and the greater that difference, the higher the confidence in the ID.

Model comparisons reported by Wixted et al. (2018) generally favored the Ensemble model over the Independent Observations model, and both of those models far outperformed another lineup signal detection model known as the Integration model (Duncan, 2006). According to the Integration model, the MAX face is identified if the *sum* of the memory signals generated by the faces in the lineup exceeds a decision criterion. Thus, although all three models assume that the MAX face is the only candidate for being identified, the Integration model is unique in that it assumes that the memory signal upon which the decision is based is not specifically tethered to the MAX face. Instead, the relevant memory signal is the sum of the memory signals generated by the faces in the lineup. Despite its intuitive implausibility, this model has long been assumed to accurately characterize lineup memory (e.g., Duncan, 2006;

Horry, Brewer, Weber & Palmer, 2015; Palmer & Brewer, 2012; Palmer, Brewer & Horry, 2013; Palmer, Brewer & Weber, 2010; Smith, Wells, Lindsay, & Penrod, 2017; Smith, Wells, Smalarz, & Lampinen, 2018). In fact, it seems fair to say that, although signal detection theory had only limited influence in the field of eyewitness identification until recently, to the extent that it did have influence, it was based on the Integration model. Yet the Integration model had never been rigorously tested. As it turns out, and perhaps not surprisingly, it does not accurately characterize lineup data and thus should probably no longer be taken seriously (Wixted et al., 2018).

Receiver operating characteristic (ROC) analysis. We are now in a position to consider how to compare the diagnostic accuracy of competing lineup procedures when both the HR and the FAR are lower in one condition compared to the other, as we saw earlier for sequential lineups compared to simultaneous lineups. How do you pick a winner in a case like that? The answer is ROC analysis. Although unknown to the world prior to 1953, signal detection theory immediately led to this groundbreaking methodological advance (Wixted, 2019). As illustrated in Figure 6, an ROC is nothing more than a plot of the HR vs. the FAR across different levels of response bias, holding discriminability constant. Various methods have been used to induce subjects to shift the decision criterion across conditions while holding discriminability constant (Swets, Tanner & Birdsall, 1955; Tanner, Swets & Green, 1956). For example, prior to presenting the lineup, instructions can be used to encourage subjects to adopt either a

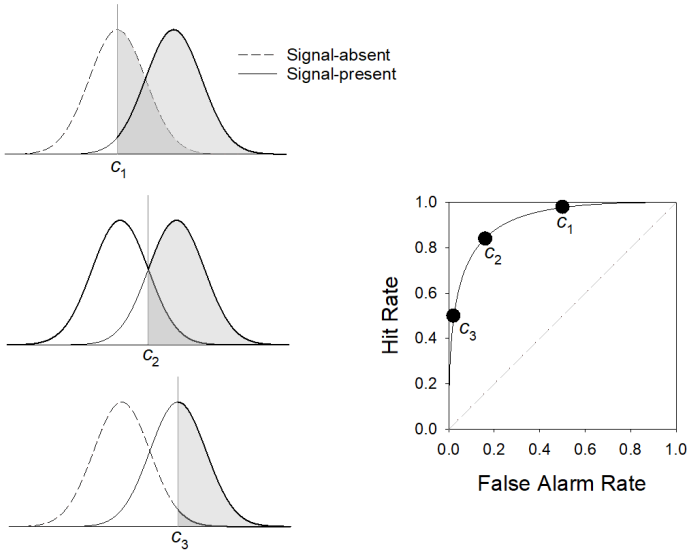

Figure 6. Signal detection interpretation of ROC data. As the criterion moves from liberal (c_1) to conservative (c_3), both the hit rate and the false alarm rate decrease (shaded regions to the right of the criterion). Note that discriminability remains constant at $d' = 2$ in this example.

conservative criterion or a liberal criterion (instruction method). In the conservative condition, the instructions might be: "Please do not make an ID unless you are sure that it was presented," whereas in the liberal condition, the instructions might instead be: "Please ID the person who is most likely to be the perpetrator even if you are not sure it is him." Each biasing condition would have a different *HR* and *FAR*, but they would (ideally) all reflect the same ability to detect the stimulus because stimulus magnitude is held constant (Mickes et al., 2017).

A simpler and arguably better way to generate ROC data is to collect confidence ratings in a single lineup condition involving a neutral response bias (the rating method). As illustrated earlier in Figure 5, positive IDs from a lineup made with varying levels of confidence can be conceptualized in exactly the same way as decisions based on criteria ranging from liberal to conservative. The most conservative (i.e., lower left) ROC point is obtained by only counting suspect IDs made with the highest level of confidence (5 in this example). The high-confidence HR might be .31 (i.e., 31% of target-present lineups yielded a correct suspect ID made with high confidence), and the high-confidence FAR would be .01 (i.e., 1% of target-absent lineups yielded

an incorrect suspect ID made with high confidence). The next (slightly more liberal) ROC point is obtained by counting suspect IDs decisions made with confidence ratings of either 4 or 5, in which case the HR might be .64 and the FAR might be .05; the next (slightly more liberal still) ROC point is obtained by counting suspect IDs decisions made with confidence ratings of 3, 4 or 5, in which case the HR might be .84 and the FAR might be .16; and so on. Continuing in this manner would yield five separate hit and false alarm rates that could be plotted on the ROC.

A critical point to understand about the ROC is that it depicts hit and false alarm rates that are *all achievable* for a given lineup procedure. Thus, instead of basing a decision about a lineup on a single pair of hit and false alarm rates, the entire family of achievable hit and false alarm rates must be considered. The reason is that a lineup procedure is not constrained to yield a single hit and false alarm rate; instead, it is constrained to the family of hit and false alarm rates that fall on the ROC curve it generates. The diagnostically superior procedure is the one that yields the higher ROC because, for any point on the lower ROC, the procedure that yields the higher ROC can achieve both a higher HR and a lower FAR. Thus, the procedure that yields the higher ROC is objectively superior to the procedure that yields the lower ROC.

With this theoretical and methodological background in mind, we now consider the three main issues that are the focus of his chapter: (1) Simultaneous vs. sequential lineups, (2) the reliability of eyewitness memory (including the confidence-accuracy relationship) and (3) the role of estimator variables in eyewitness identification.

Simultaneous vs. Sequential Lineups

As noted earlier, a simultaneous photo lineup involves the simultaneous presentation of all of the faces in the lineup. In a sequential lineup, the photos are instead presented one at a time (Lindsay & Wells, 1985; Steblay, Dysart, & Wells, 2011). In most experimental studies of the

sequential lineup, a stopping rule is used such that the first photo that elicits a "yes" response terminates the procedure. The idea that sequential lineups might be superior to simultaneous lineups was based primarily on the results of mock-crime laboratory experiments conducted over the last 30 years. To determine which lineup procedure is superior in an applied sense, the focus has always been primarily placed on consequential suspect IDs (i.e., on the HR and FAR), not on filler IDs. However, until fairly recently, the HR and FAR were not conceptualized in terms of signal detection theory and ROC analysis, and that is what led the field astray.

Measures of Diagnostic Accuracy

The diagnosticity ratio. The original argument in favor of the sequential lineup procedure comes from combining the correct and incorrect suspect ID rates into a ratio known as the diagnosticity ratio (DR). More specifically, DR = HR / FAR. The DR is what is usually thought of as the likelihood ratio in the odds version of Bayes' theorem, according to which the posterior odds of guilt are equal to the prior odds of guilt multiplied by the likelihood ratio. A single DR can be computed using all correct and false IDs (as it invariably is when used to compare the diagnostic accuracy of simultaneous vs. sequential lineups) or separately for each level of confidence (as is often done when measuring the confidence-accuracy relationship). More formally, Bayes' theorem compares the odds in favor of one hypothesis over another. The two hypotheses of interest here are:

H_1: the suspect is guilty

H_2: the suspect is innocent

Bayes' theorem states that:

$$\frac{P(H_1|D)}{P(H_2|D)} = \frac{P(D|H_1)}{P(D|H_{2)}} \frac{P(H_1)}{P(H_2)}$$

where D is the data (a suspect ID in this case), $P(H_1|D) / P(H_2|D)$ represents the posterior odds of H_1 compared to H_2 (i.e., the odds of guilt after a suspect ID has been made), $P(D|H_1) / P(D|H_2)$ represents the likelihood ratio (i.e., the diagnosticity ratio), and $P(H_1) / P(H_2)$ represents the prior odds of H_1 compared to H_2 (i.e., the odds of guilt before a suspect ID has been made).

In the lineup scenario, $P(D|H_1)$ is the HR (i.e., the correct ID rate), and $P(D|H_2)$ is the FAR (i.e., the false ID rate). Thus, the DR (i.e., the likelihood ratio) is equal to the correct ID rate divided by the false ID rate (i.e., HR / FAR). In most experiments, half the participants are presented with a target-present lineup and half are presented with a target-absent lineup, which means that the base rate of guilt equals the base rate of innocence. Under such conditions, the prior odds of guilt, $P(H_1) / P(H_2)$, equal 1, in which case the DR directly indicates the posterior odds of guilt. For example, if the prior odds are equal to 1, and if the HR = .50 and the FAR = .10, the resulting DR of 5 would mean that a suspect identified using this procedure is 5 times as likely to be guilty as innocent.

Note that this measure is computed from witnesses who identify a suspect, excluding witnesses who pick a filler or make no ID. The reason is that the information value of the three possible lineup decision outcomes differ. A suspect ID points in the direction of guilt, but filler IDs and No IDs both point in the direction of innocence. Thus, it is essential to compute the posterior odds of guilt separately for each decision outcome (Wells, Yang, & Smalarz, 2015). A suspect ID is the only outcome that imperils anyone in the lineup, so we focus on that outcome here. Later, we describe how the DR computed from suspect IDs has been used to assess the confidence-accuracy relationship, but here we focus on its use in comparing the diagnostic accuracy of different lineup formats.

As noted earlier, Lindsay and Wells (1985) reported that for the sequential lineup, HR = .50 and FAR = .17 (DR_{SEQ} = .50 / .17 = 2.94), whereas for the simultaneous lineup, HR = .58 and FAR = .42 (DR_{SIM} = .58 / .42 = 1.38). Steblay et al. (2011) later reviewed the relevant literature, arguing that this empirical pattern is representative of later studies. It seems fair to say that, to many, the large reduction in the FAR is what makes the sequential procedure so attractive. However, upon reflection, it becomes clear that one must consider the effect on the HR as well. At first, the relatively small decrease in the HR associated with switching to the sequential procedure seems reassuring. However, because this "hand waving" analysis of the effect of sequential lineups on the HR and FAR is clearly insufficient, a quantitative assessment of some kind is needed. The DR has often by used for this purpose.

Unfortunately, the DR computed from the overall HR and FAR does not provide the information needed to decide whether simultaneous lineups or sequential lineups better enable eyewitnesses to distinguish innocent from guilty suspects (i.e., it is not a measure of discriminability). As noted above, the DR instead quantifies the posterior odds of guilt (which is relevant to measuring the confidence-accuracy relationship), not discriminability. Indeed, as we illustrate next, there is only one measure of discriminability for a given lineup procedure, but each lineup procedure is associated with many different HRs and FARs and, therefore, many different DRs.

This now well-known problem with the DR is illustrated in Figure 7 (Wixted & Mickes, 2018). Note that the rightmost

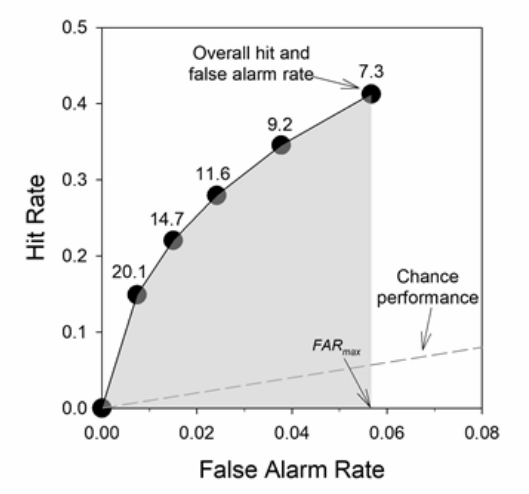

Figure 7. Hypothetical receiver operating characteristics (ROC) curve for a lineup procedure in which a 5-point confidence scale was used. The rightmost ROC point represents the overall correct and false positive identification (ID) rates that are ordinarily used to compute the diagnosticity ratio, but it can be computed for each point on the ROC. The number above each point is the diagnosticity ratio for that correct and false identification (ID) rate pair. The region shaded in light gray represents the partial area under the ROC curve (pAUC) for the specified false ID rate range of 0 to FAR max, which is equal to .057 in this case. The diagonal line represents chance performance (where correct ID rate = false ID rate).

ROC point is what most people think of as the HR and FAR. That is, it represents the HR and FAR with every ID counted, regardless of confidence. In non-lineup tasks, the FAR typically ranges from 0 to 1, but in a fair 6-person lineup ROC, it only ranges from 0 to 1/6 (.167). That is the maximum possible FAR because even if every witness presented with a fair target-absent lineup identified someone (i.e., if responding were maximally liberal such that a "yes" response were made to every target-absent lineup), witnesses would land on the innocent suspect by chance only 1/6 of the time and would land on a filler the other 5/6 of the time. In actual practice, the obtained FAR is typically much less than .167 (as it is in Figure 7) because, typically, responding is not maximally liberal. For the overall HR and FAR in Figure 7 (i.e., for the rightmost ROC point), DR = 7.3. This is the DR that is typically used to characterize the diagnostic accuracy of a lineup, but it has no special status relative to the other DRs that the

same lineup can also achieve. Standing alone, no one of the five DRs in Figure 7 tells you how far above the diagonal line of chance performance the ROC curve falls.

Partial area under the ROC. A much better measure of diagnostic accuracy—the one that flows from signal detection theory—is the area under the ROC. Because the FAR for a lineup is limited to a range that is less than 0 to 1, the relevant measure of empirical discriminability for a lineup is the *partial* area under the curve (Mickes, Flowe & Wixted, 2012; Gronlund, Wixted, & Mickes, 2014). The partial area under the curve (pAUC) is computed from a false alarm rate of 0 up to some maximum that is less than or equal to .167. That maximum FAR is denoted here as FAR_{max}. An obvious choice for FAR_{max} is the FAR associated with the overall hit and false alarm rate that corresponds to the rightmost ROC point (.413 and .057, respectively, in Figure 7). With $FAR_{max} = .057$, as it is in Figure 7, pAUC for these data (i.e., the area of the shaded region) is approximately .017.

What does the pAUC measure actually tell you? On its own, not very much. However, the usual goal is to compare the pAUC values for two different lineup procedures. That comparative analysis is extremely informative because the procedure that yields the higher pAUC is the diagnostically superior procedure over the false alarm rate range used in the analysis. To compare the two procedures with respect to pAUC, it is essential to use the same FAR_{max} to measure the area under both curves. One reasonable strategy is to set FAR_{max} equal to the FAR associated with the rightmost point of the more conservative of the two procedures being compared, as illustrated using hypothetical data in Figure 8 (Wixted & Mickes, 2018). Such an analysis covers a range that includes empirical ROC data generated by both procedures

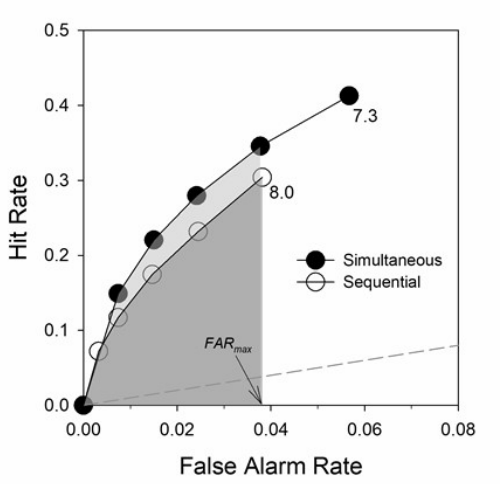

Figure 8. Hypothetical receiver operating characteristic (ROC) curves for two eyewitness identification procedures in which a 5-point confidence scale was used. Note that the diagnosticity ratio for the rightmost point is higher for the sequential procedure, a result that was once interpreted to mean that the sequential procedure is diagnostically superior to the simultaneous procedure. The region shaded dark gray represents the partial area under the curve (pAUC) for the sequential procedure in the specified false ID rate range of 0 to FAR_{max}. That dark gray region plus the light gray region above it represents the pAUC for the simultaneous procedure over the same false ID rate range.

and therefore does not involve theoretically extrapolating the ROC curve to the right for either procedure. In Figure 8, it is visually obvious that pAUC for the simultaneous procedure is greater than the pAUC for the sequential procedure over the FAR range of 0 to .038 (the maximum FAR for the sequential procedure). This is true even though, as ordinarily computed, the DR for the sequential procedure (8.0) is larger than the DR for the simultaneous procedure (7.3). In the past, such an outcome would have been interpreted as reflecting a sequential superiority effect, but this example actually illustrates a simultaneous superiority effect.

Since 2011, which is the last year that anyone claimed a sequential superiority effect (Steblay et al., 2011), a number of studies have found a *simultaneous* superiority effect such that $pAUC_{SIM} > pAUC_{SEQ}$. We gathered and analyzed all published and unpublished ROC-based comparisons known to us (17 in all), computing an effect-size measure (Cohen's *d*) for each. The results are shown in Figure 9. Of the 17 comparisons, 15 favor the simultaneous procedure, 9 significantly, and 2 favor the sequential procedure, both non-significantly. Moreover, two police

department field studies have reported findings consistent with the results of these lab studies, favoring the simultaneous lineup procedure (Amendola & Wixted, 2015a,

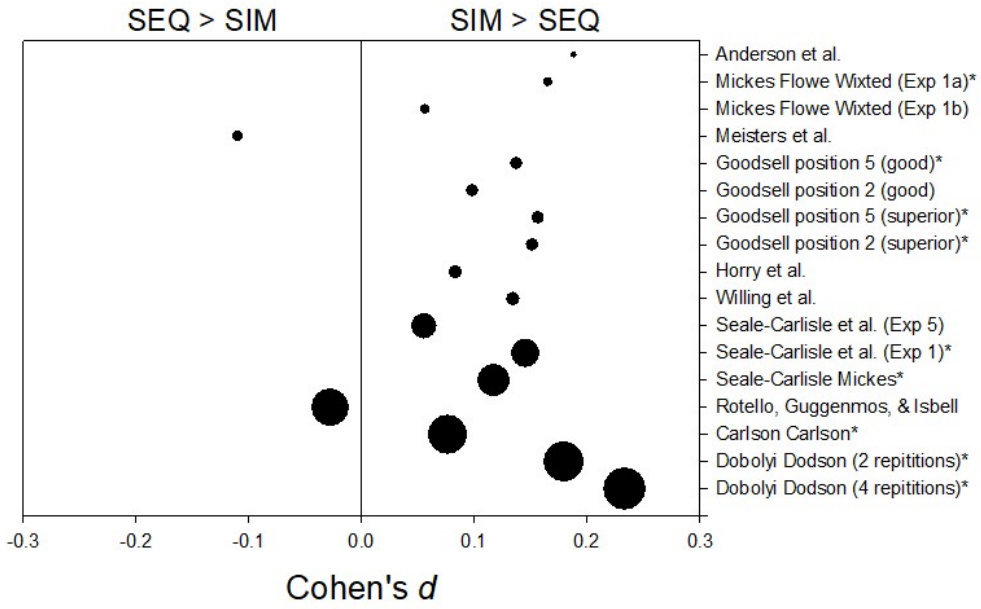

Figure 9. Estimated Cohen's *d* scores from 17 published and unpublished ROC-based comparisons of simultaneous (SIM) vs. sequential (SEQ) lineups. The scores were estimated using the formula $\text{Cohen's } d = 2D/\sqrt{N}$, where *D* is the test statistic returned by the pROC software (Robin et al., 2011). The data are arranged from smallest *N* to largest *N* (top to bottom) as indicated by symbol size, and the significant results are indicated by an asterisk next to the study reference. Note that, as used here, *N* refers to the number of lineups tested (which, in most cases equals the number of subjects tested). Some of the SEQ procedures (namely, the 3 comparisons with Seale-Carlisle as an author) were like those used in the UK, which involves a sequential "video parade." Asterisks on the citations indicate a significant difference ($p < .05$).

2017; Wixted, Mickes, Dunn, Clark, & Wells, W., 2016). Although there was some debate about one of these studies (Amendola & Wixted, 2015b, 2017; Wells, Dysart, & Steblay, 2015), based on the available evidence, it seems fair to reject the notion of a "sequential superiority effect." Instead, the data indicate a discriminability advantage for simultaneous lineups (i.e., $pAUC_{SIM} > pAUC_{SEQ}$). Note that on the basis of research that erroneously relied on the DR instead of pAUC to measure the diagnostic accuracy of different lineup procedures between 1985 and 2011, approximately 30% of more than 17,000 law enforcement agencies in the U.S. adopted the sequential procedure (Police Executive Research Forum, 2013). In terms of applied impact, this work falls among the most influential lines of research in the history of psychology. The story of

how so many came to believe that sequential lineups are diagnostically superior to simultaneous lineups illustrates the danger of analyzing recognition memory data without relying on a model like signal detection theory or related models (Clark, 2003).

Recently proposed alternative measures. For the most part, eyewitness ID researchers have been interested in determining which procedure yields superior diagnostic performance over a range in which even FAR_{max} is low, thereby keeping the risk of falsely identifying an innocent suspect low. Recently, however, researchers have argued that pAUC is a problematic measure if maintaining a low FAR is no longer a priority (Smith et al., 2018). Their concern is a surprising one because it is based on the idea that 6-person lineups, with the maximum FAR constrained to be 1/6 = .167, may be overly protective of innocent suspects. To take an extreme example to illustrate the argument, imagine that the police convince themselves that out of every 1000 lineups they perform, 999 contain a guilty suspect and only 1 contains an innocent suspect (i.e., the base rate of guilt is assumed to be extremely high). In that case, a procedure that yields a high false ID rate might be acceptable (even desired) in order to achieve the higher HR that would come with it. A high HR would properly imperil many guilty suspects, but a high FAR would imperil very few innocent suspects because, at least in the minds of the police, lineups rarely contain an innocent suspect.

A showup coupled with very liberal responding can achieve a FAR all the way up to 1.0, but the FAR for a fair 6-person lineup hits a ceiling of .167 even for maximally liberal responding. Because an analysis using pAUC requires a common FAR range for both procedures being compared (e.g., a lineup vs. a showup), it cannot identify the superior procedure over a range in which $FAR_{max} > .167$. To some, this limitation disqualifies pAUC as a general measure

of the diagnostic accuracy of competing procedures. Instead of using pAUC, Smith et al. (2018) have argued in favor of a new measure known as "deviation from perfect performance" (DPP).

"Perfect" performance is a maximum-utility outcome defined by (1) subjective assumptions about the base rate of guilty suspects in eyewitness identification procedures, and (2) subjective values associated with different outcomes (e.g., the tradeoff in costs associated with misidentifications of the innocent and failures to identify the guilty). A utility calculation might show that "perfect" (i.e., maximum utility) performance can be achieved using the showup instead of the lineup, even if the showup yields a lower pAUC up to a FAR of .167, as is typically the case (e.g., Neuschatz, Wetmore, Gronlund & Goodsell, 2016; Wetmore et al., 2015). For example, a showup with a HR of .50 and a FAR of .33 might be associated with higher utility than any HR-FAR achievable using a lineup. In that case, pAUC could not identify the superior procedure, but DPP could.

Unfortunately, DPP has a fatal flaw, namely, its reliance on both subjective assumptions and subjective values, in which case diagnostic superiority is in the eye of the beholder, not in the objective memory performance of eyewitnesses. The whole reason an identification procedure is administered to an eyewitness is to find out if, based on the memory of the eyewitness (and nothing else), the police have the right suspect. Yet in no sense is DPP a measure of eyewitness memory because its value is so heavily dependent on subjective assumptions the police make about base rates and on subjective costs assigned to the two main kinds of errors eyewitnesses can make (namely, misidentifying the innocent suspect and failing to identify the guilty suspect). Thus, what is needed is an independent measure of memory performance irrespective of subjective beliefs about base rates and subjective values, and that is

exactly what pAUC provides. Subjective considerations can be factored in later, when deciding where the decision criterion should be placed to achieve the desired HR and FAR.

What, then, should one do if a higher FAR than can be achieved using a standard lineup is actually judged to be a desirable outcome (an unlikely state of affairs, to be sure)? That is, which procedure would be superior, the lineup (max FAR = .167) or the showup (max FAR = 1.0)? Instead of abandoning objective pAUC and placing the answer into the eye of the beholder, we suggest that a better solution would be to simply extend the lineup ROC to the right (i.e., beyond the usual FAR limit of .167). At first glance, it seems impossible, but it is actually easy to do. If the protection of innocent suspects is of reduced concern because of the subjective assumption that lineups almost always contain a guilty suspect, then keeping the eyewitness blind to the suspect in the lineup would no longer be a priority. If the witness knows who the suspect is, then maximally liberal responding will now yield a FAR of 1.0 rather than .167. Colloff and Wixted (2019) showed that by simply unblinding the simultaneous lineup procedure in this way (i.e., by informing the witness who the suspect in the lineup is), the ROC can be extended to the right. Critically, as measured by pAUC, the non-blind lineup still retains higher discriminability than the equally non-blind showup. Thus, using this simple approach, one can achieve both the desired higher FAR *and* the higher discriminability afforded by a simultaneous lineup procedure without abandoning a purely objective measure of memory performance (namely, pAUC).

Theories of Lineup Performance

Absolute vs. Relative Judgments. When the sequential procedure was believed to be superior to the simultaneous procedure, the dominant theoretical explanation was based on the distinction between absolute vs. relative judgments. More specifically, Wells (1984) proposed

that simultaneous lineups create a tendency to identify the lineup member who most resembles the eyewitness's memory of the perpetrator (a *relative* decision strategy). In the extreme case, the use of this decision strategy would lead eyewitnesses to always choose someone from a lineup, which is problematic in the case of target-absent lineups because of the high number of false identifications that would occur. An alternative strategy is an *absolute* decision strategy, which Lindsay and Wells (1985) argued can be promoted by using a sequential lineup. Instead of choosing the lineup member who looks most like the perpetrator, eyewitnesses who use this strategy would evaluate each lineup member against an absolute decision criterion. If no one in the lineup yielded a strong enough match to the eyewitness's memory of the perpetrator (i.e., a strong enough match to exceed the decision criterion), the lineup would be rejected. Because the tendency to identify someone from a lineup is lower when an absolute strategy is used, the likelihood of misidentifying an innocent suspect (i.e., the false ID rate) is correspondingly lower.

Why does the use of an absolute strategy increase the DR associated with the overall HR and FAR? Because an absolute strategy is another term for a conservative response bias, and as illustrated earlier in Figure 7, adopting a more conservative response bias would result in an increase in the DR. Indeed, in his original article on the topic, Wells (1984) wrote, "It is possible to construe of the relative judgments process as one that yields a response bias, specifically a bias to choose someone from the lineup" (p. 94). Thus, a relative judgment strategy is just another term for a liberal response bias. The implication is that the absolute/relative theory is a theory about response bias, and it might even be correct. For example, because sequential lineups yield conservative responding, and because the DR increases as responding becomes more conservative (Figure 7), it correctly predicts that $DR_{SEQ} > DR_{SIM}$. However, it says nothing about

which procedure is diagnostically superior to the other, so it does not explain why simultaneous lineups usually yield higher discriminability than sequential lineups.

The absolute/relative judgment theory was not grounded in an established cognitive principle of discrimination or decision-making but was instead a novel perspective. Some attempts have been made to formalize it in models that are compatible with signal detection theory (Clark et al., 2011; Fife et al., 2014; Moreland & Clark, 2019), but those studies have not supported the notion that absolute decisions yield a diagnostically superior outcome compared to relative decisions. An alternative account proposed to explain the unexpected simultaneous superiority effect that became apparent with the advent of ROC analysis is known as diagnostic feature-detection theory (Wixted & Mickes, 2014). As described next, this theory is grounded in longstanding principles of perceptual learning and decision-making.

Diagnostic Feature-Detection Theory. Gibson (1969) reviewed a large body of evidence from the perceptual learning literature and identified a key principle. Specifically, she concluded that an important step in perceptual learning—that is, learning to discriminate similar objects—involves the detection of distinctive (i.e., diagnostic) features. In perceptual learning experiments, such as learning to tell the difference between two breeds of dog, it has often been found that presenting similar objects simultaneously facilitates the detection of the distinctive features that serve to differentiate the objects, thereby enhancing discriminability between them, compared to when the objects are presented sequentially (Gibson, 1969). Theoretically, the simultaneous presentation of two similar exemplars makes it easier to identify their distinctive features (e.g., Collies have longer hair than German Shepherds) and to ignore their many common features (e.g., both have four legs, a tail, etc.) compared to when the exemplars are presented sequentially.

Building upon these ideas and related ideas proposed by Tversky (1977), Wixted and Mickes (2014) proposed that a similar process involving the detection of distinctive (and, therefore, diagnostic) features plays an important role at the time an eyewitness identification is made. Consider, for example, an eyewitness who sees a White male in his early 20s rob a liquor store. The witness might describe the perpetrator to the police using those terms (i.e., White male in his early 20s), but the witness's memory representation of the perpetrator might also include additional unmentioned details, such as the fact that the perpetrator has an oval face and small eyes. Later, the police might identify a suspect who matches the general description provided by the eyewitness (viz., a White male in his early 20s). In a 6-person lineup, all six members will match the general description of the suspect— that is, the suspect (innocent or guilty) and the fillers will all be young White males. Critically, what this means is that the features used to qualify suspects and fillers for inclusion in a lineup are non-diagnostic of guilt. Yet, if not discounted, these features will generate noisy memory-match signals that will sometimes be randomly strong. Thus, the more weight the witness attaches to those features, the harder it will be for the witness to distinguish innocent from guilty suspects (who, by design, will share those features). For example, if those were the *only* features taken into consideration, then d' would be 0.

When a face is presented to the eyewitness in isolation as part of a sequential lineup (or in a showup), there is no obvious indication that some features are non-diagnostic in terms of distinguishing between innocent and guilty suspects. To the extent that the non-diagnostic features are given weight under those circumstances, the ability to discriminate innocent from guilty suspects will suffer. In a simultaneous lineup, by contrast, it is immediately apparent to the eyewitness that everyone in the lineup shares certain non-diagnostic features (i.e., it is

immediately apparent that they are all young White males). For that reason, the eyewitness will be induced to attach weight to features that might be diagnostic (e.g., shape of face and size of eyes) while discounting the nondiagnostic features. To the extent that they do, the ability to discriminate an innocent suspect from a guilty suspect will be enhanced, thereby elevating the ROC.[1] Thus, diagnostic feature-detection theory is not a theory of response bias, as the absolute/relative judgment theory is; instead, it is a theory of discriminability.

Evidence consistent with the diagnostic feature-detection theory is now abundant. In addition to accounting for the simultaneous superiority effect (relative to sequential lineups), it also naturally accounts for the fact that the simultaneous ROC is higher than the showup ROC (Wetmore et al., 2015; Neuschatz et al., 2016). Indeed, as noted earlier, Colloff and Wixted (2019) found that even when a box was placed around the suspect in a simultaneous lineup – making it a "suggestive" procedure (every bit as suggestive as a showup) – discriminability still exceeded that of the showup. The only effect of making the suspect known to the eyewitnesses in this way was to increase the overall HR and FAR, extending the simultaneous ROC to the right beyond its usual limit of .167.

Colloff et al. (2016) tested another prediction made by the diagnostic feature-detection theory. Specifically, they tested whether an unfair lineup (defined as a lineup in which the suspect – innocent or guilty – matches the memory of the perpetrator better than the fillers do), would reduce discriminability. The perpetrator who was seen committing a simulated crime in their study had a black eye. In the fair lineups, everyone in the lineup had a black eye or no one in the lineup did (including the guilty suspect). In the unfair lineup, by contrast, only the suspect

[1] Before it was known that simultaneous lineups are diagnostically superior to sequential lineups, Goodsell, Gronlund, and Carlson (2010) proposed the same idea to account for the possibility that, in a sequential lineup, discriminability is higher when the suspect is placed in a late position compared to an early position.

(innocent or guilty) had a black eye. In other words, the black eye feature was not shared by the fillers yet it did match the eyewitness's memory of the perpetrator. Diagnostic feature-detection theory therefore predicts that it will not be discounted. Because that feature does not distinguish innocent from guilty suspects in unfair lineups, discriminability should suffer accordingly (i.e., the unfair ROC should be lower than the fair ROC). As predicted, the pAUC was lower for unfair lineups compared to fair lineups (Colloff et al., 2016).

Filler siphoning theory. Another theory that has been advanced to explain why simultaneous lineups yield higher discriminability than both showups and unfair lineups (yet is silent about why simultaneous lineups yield higher discriminability than sequential lineups) is "filler-siphoning theory" (e.g., Smith et al., 2018). According to this account, the presence of similar-looking ("attractive") fillers in fair lineups siphon away more identifications from the innocent suspect in target-absent lineups than from the guilty suspect in target-present lineups. In other words, compared to unfair lineups (where the fillers play a reduced role) or showups (where fillers play no role at all), the differential filler siphoning process in fair lineups reduces the FAR more than the HR. This is another way of saying that filler siphoning increases the DR, which is surely true (Clark, 2012; Fitzgerald, Price, Oriet, & Charman, 2013). It is also the same argument once advanced in favor of the so-called sequential superiority effect. As hard as it might be to believe, reducing the HR more than the FAR (i.e., increasing the DR) does not automatically reflect a diagnostically superior state of affairs (see Figure 8). The fact that the same idea could arise in a different context (now, comparing fair lineups to unfair lineups and showups) long after the notion of a sequential superiority effect has been abandoned illustrates just how powerful this erroneous intuition can be. Indeed, the idea that reducing the FAR more

than the HR constitutes a superior state of affairs is almost impossible to let go of until one understands signal detection theory and ROC analysis.

As noted by Colloff, Wade, Strange and Wixted (2018), filler siphoning is a behavioral phenomenon, not a theory. Moreover, it is a behavioral phenomenon that is inherently predicted by any signal detection model. According to signal detection theory, adding attractive fillers to a lineup will siphon away IDs that would otherwise land on the innocent or guilty suspect, and it will do so in a way that reduces the FAR more than the HR, just as using a more conservative criterion would. However—and this is the key point—that would happen if the ability to discriminate innocent from guilty suspects increased, decreased or remained the same as a result of adding those attractive fillers. As it turns out, and as uniquely predicted by diagnostic feature-detection theory, adding attractive fillers (i.e., increasing lineup fairness by, for example, placing a black eye on everyone in the lineup) increased the ability of eyewitnesses to discriminate innocent from guilty suspects. Theoretically, discriminability increased because, now, the black-eye feature was discounted, allowing the eyewitness to base the decision on facial features that distinguish innocent from guilty suspects.

Two other findings weigh against the notion that filler siphoning provides an explanation of why simultaneous lineups yield higher discriminability than showups. First, as noted above, Colloff and Wixted (2019) found that by simply identifying the suspect in the simultaneous lineup to the eyewitness, the ROC was extended to the right (beyond its usual FAR limit of .167) while remaining elevated compared to a showup. Although fillers were present in the unblinded simultaneous lineup, no filler siphoning occurred (i.e., no subject picked a filler because the suspect was highlighted), so that cannot be the explanation for why discriminability was enhanced by the simultaneous presentation of faces. However, diagnostic feature-detection

theory accounts for that finding because the presence of the fillers in the simultaneous showup theoretically induces witnesses to discount non-diagnostic facial features (just as in a standard simultaneous lineup).

Second, Colloff et al. (2018) compared two *showup* conditions in which diagnostic feature-detection theory predicts a discriminability difference, one that could not possibly be explained by filler siphoning because no fillers are present in either condition. This experiment was exactly like the fair-vs.-unfair lineup comparison reported by Colloff et al. (2016) except that the fillers were eliminated. After watching a video of a perpetrator with a black eye, participants in the "unfair" showup condition were presented with either an innocent suspect with a black eye or a guilty suspect with a black eye. As in the case of unfair lineups, this distinctive feature should not be discounted (i.e., the black eye will be given weight), and because it matches the memory trace but does not distinguish innocent from guilty suspects, discriminability should be harmed accordingly. In the "fair" showup condition, by contrast, *neither* suspect had the distinctive feature because the area of the feature was covered with a black box. Diagnostic feature-detection theory makes the same prediction as in the corresponding fair-vs.-unfair lineup comparison. That is, theoretically, a fair showup prevents witnesses from relying on a nondiagnostic feature by removing it altogether from the decision-making process (i.e., removing a noisy, non-diagnostic memory signal), enhancing the ability of witnesses to discriminate between innocent and guilty suspects. Analogously, in fair lineups, similar foils who share the distinctive feature effectively remove it by causing that feature to be discounted, again enhancing the ability of witnesses to discriminate between innocent and guilty suspects. As predicted by diagnostic feature-detection theory (but not by filler-siphoning theory), pAUC for the fair showup was greater than pAUC for the unfair showup.

Criterion variability theory. Yet another theory recently proposed by Smith et al. (2017) holds that showups yield the same underlying discriminability as simultaneous lineups (i.e., $d'_{SIM} = d'_{Showup}$), but their empirical ROCs measured in terms of pAUC (i.e., $pAUC_{SIM} > pAUC_{Showup}$) differ due to other factors. Specifically, according to this model, different eyewitness identification procedures are differentially susceptible to the deleterious effects of criterion variability. Criterion variability refers to the uncontroversial fact that different participants will place the criterion in different locations (i.e., some will be conservative, others more liberal). Their simulations showed that, in the absence of criterion variability and with underlying d' equated, lineups and showups produce comparable empirical ROC curves up to a FAR of 0.167 (and, therefore, comparable pAUC values over that range). However, in the presence of criterion variability (equated across the two procedures), simultaneous lineups yielded higher empirical discriminability (measured by pAUC) than showups despite underlying d' being equated for the two procedures. Because the existence of criterion variability is non-controversial, Smith et al. interpreted their simulation results as a demonstration that underlying discriminability is in fact equated for the two procedures, contrary to diagnostic feature-detection theory. In truth, their simulation is merely an existence proof, one that puts a new theory on the table in addition to diagnostic feature-detection theory.

Criterion variability theory was not applied to simultaneous vs. sequential lineups. Using simulations, Wixted and Mickes (2018) confirmed that given equal d' and equal criterion variability, the simultaneous ROC would be higher than the showup ROC (i.e., pAUC would be higher for the simultaneous lineup procedure). However, they also extended the predictions of this model by showing that criterion variability theory further predicts that the simultaneous lineup ROC should be higher than the sequential lineup ROC. Thus, there are now two

theoretical reasons to expect simultaneous lineups to outperform both sequential lineups and showups, diagnostic feature detection theory and criterion variability theory. The fact that two different theories predict a simultaneous superiority effect reinforces the conclusion from recent empirical research (summarized earlier in Figure 9) that consistently points in the same direction.

Although diagnostic feature-detection theory and criterion-variability theory could both be true, the available evidence is more supportive of the former theory. For example, with regard to two procedures we discussed earlier, criterion-variability theory makes no a priori prediction about fair vs. unfair *showups* (Colloff et al., 2018) or between simultaneous showups (i.e., unblinded lineups in which the suspect is identified) vs. standard showups (Colloff & Wixted, 2019). However, diagnostic feature-detection theory does. As predicted by that theory, fair showups yielded a higher ROC than unfair showups, and simultaneous showups yielded a higher ROC than standard showups (see also Wetmore, McAdoo, Gronlund, & Neuschatz, 2017).

The Reliability of Eyewitness Memory

To this point, we have considered the diagnostic accuracy of different eyewitness identification procedures. However, an issue of even wider interest in the field concerns the reliability of eyewitness memory in general. Ask any class of students if they believe eyewitness memory is (a) reliable or (b) unreliable and virtually everyone will choose the latter option. One can hardly blame them. After all, as noted earlier, the case in favor of the idea that eyewitness memory is unreliable, regardless of confidence, seems strong. First, convincing research shows that memory is malleable, so much so that people can come to confidently remember traumatic events that never actually happened (e.g., Loftus, & Ketcham, 1994). Second, lab-based research was long interpreted to mean that the confidence an eyewitness expresses upon identifying someone from a lineup is not particularly indicative of accuracy, not even under “pristine”

testing conditions (e.g., Penrod & Cutler, 1995; Wells & Murray, 1984). Even then, highly confident eyewitnesses are often wrong. Third, and most compelling of all, eyewitness misidentifications made with high confidence in a court of law are known to have played a role in more than 70% of the more than 360 wrongful convictions that have been overturned based on DNA evidence since 1989 (Innocence Project, 2019). On the surface, eyewitness memory appears to be unreliable no matter how confident the eyewitness might be.

Against this almost universal perspective, we have argued that eyewitness memory is highly reliable on the first test conducted early in a police investigation (Wixted, Mickes, & Fisher, 2018). This is true of both recall (namely, a properly conducted police interview) and recognition (namely, a properly conducted police lineup). Critically, the very act of testing recognition memory using a lineup or a showup contaminates it in the sense that the suspect, innocent or guilty, will be more familiar to the witness on any later test. In addition, one's recall of details can be contaminated by misinformation (e.g., from inaccurate news stories covering a witnessed crime). Thus, the reliability of eyewitness memory is never higher than it is when first tested – and never lower than it is when ultimately tested in a court of law in front of a jury.

Measuring the Confidence-Accuracy Relationship

The question of whether eyewitness memory is reliable is a question about the relationship between confidence and accuracy. Eyewitness memory would be reliable if high confidence implied high-accuracy and low confidence implied low accuracy. However, eyewitness memory would be unreliable if accuracy were low across the board, regardless of confidence. For decades, the relevant evidence was interpreted to mean that eyewitness IDs were error-prone and that the confidence-accuracy relationship was weak, indicating that eyewitness memory is unreliable. Recently, that story has changed as well.

The correlation coefficient. A key mistake in prior lab-based research on the usefulness of eyewitness confidence on an initial test was its reliance on the correlation coefficient to quantify the confidence-accuracy relationship. However, just as the DR was the wrong way to measure the diagnostic accuracy of lineups, the correlation coefficient, despite its intuitive appeal, is the wrong way to measure the confidence-accuracy relationship. A detailed explanation of what is wrong with the correlation coefficient is presented in Appendix A of Wixted and Wells (2017). Briefly, the problem is that the data entered into the correlational analysis consists of a binary variable (e.g., 1 = correct ID vs. 0 = incorrect ID) and a continuous variable (e.g., measured using a confidence scale ranging from 1 to 100). Thus, for example, Eyewitness 1 might yield values of 1 and 70 (correct ID with a confidence rating of 70), Eyewitness 2 might yield values of 0 and 30 (incorrect ID with a confidence rating of 30), and so on. Correlating a dichotomous variable with a continuous variable yields a point-biserial correlation coefficient, and the problem is that its value will be far less than 1 even when confidence is as strongly related to accuracy as it could possibly be (Juslin, Olsson, & Winman, 1996). Thus, another approach to measuring the relationship of interest is needed.

Calibration. A much better way to measure the confidence-accuracy relationship is to plot a calibration curve (Juslin et al., 1996). Using this approach, a confidence rating of 0% means there is no chance that the identified suspect is the perpetrator, whereas a confidence rating of 100% means that the identified suspect is, beyond any doubt, the perpetrator. To construct a calibration plot, an accuracy score is computed separately for each level of confidence, and the results are plotted on a graph. For example, a high-confidence accuracy score can be computed using IDs made with confidence in the 90%-100% range, with the measured being correct IDs / (correct IDs + incorrect IDs).

When the calibration approach is used, it becomes clear that the confidence-accuracy relationship is much stronger than is suggested by computing a correlation coefficient (e.g., Brewer, Keast, & Rishworth, 2002; Brewer & Wells, 2006; Juslin et al., 1996; Weber & Brewer, 2004). However, whereas correct IDs in the calibration equation always refer to guilty suspect IDs, incorrect IDs typically consist of innocent suspect IDs and filler IDs alike. Yet it is a mistake to include both types of errors because the question of interest to a court of law concerns the reliability of a *suspect ID*. That is, at trial, the judge and jury want to know the following: given that a suspect ID has been made with a certain level of confidence, how likely is it to be correct? The inclusion of filler IDs in the denominator means that a calibration analysis, as typically performed, cannot answer that question.

Confidence-accuracy characteristic analysis. How should the relationship be measured instead? Signal detection theory provides a guide and a prediction. For guidance, imagine a list-memory task in which confidence ratings are collected using a standard 6-point scale, where 1 = "Sure No," 2 = "Probably No," 3 = "Maybe No," 4 = "Maybe Yes," 5 = "Probably Yes," and 6 = "Sure Yes." The signal detection interpretation of this scenario is illustrated in Figure 10. Although the data of interest are the same data used to plot a confidence-based ROC, when the goal is to measure the confidence-accuracy relationship instead of discriminability, the model in Figure 10 calls for a different way to analyze the data.

For an Old/New recognition memory task, confidence-specific accuracy for "yes" decisions is equal to nH_c / (nH_c + nFA_c), where nH_c is the number of hits made with a particular level of accuracy and nFA_c is the number of false alarms made with that same level of accuracy. The most informative way to characterize the confidence accuracy relationship is to simply plot

this measure of suspect ID accuracy as a function of confidence. Mickes (2015) referred to this kind of plot as a confidence-accuracy characteristic (CAC) plot.

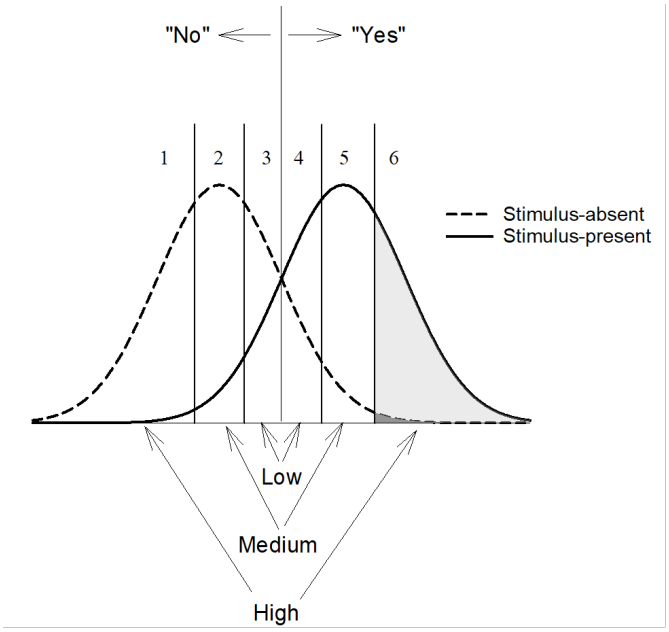

Figure 10. The signal detection interpretation of confidence ratings made using a 6-point scale (ranging from 1 = high-confidence "No" to 6 = high-confidence "Yes"). As illustrated by the shaded regions, the area under the target distribution to the right of the highest confidence criterion is .31, whereas the corresponding area under the lure distribution is .01. Thus, the high-confidence hit rate would be .31 (i.e., of all test trials involving a signal, 31% received a "yes" decision with a rating of 6), and the high-confidence false alarm rate would be .01 (i.e., of all test trials involving noise, 1% received a "yes" decision with a rating of 6). If the number of stimulus-present and stimulus-absent trials are the same (as is typically true), then the accuracy of high-confidence "Yes" decisions would be .31 / (.31 + .01) = .98 correct.

In the typical equal-base-rate scenario (i.e., the number of targets = the number of lures), the value plotted on the CAC is also known as positive predictive value (PPV). PPV varies with the base rate of old items. Thus, holding all else constant, PPV will increase as the base rate of old items increases.

To create a CAC plot that it is independent of base rate (as an ROC plot is), the relevant measure of suspect ID accuracy equal to HR_c / (HR_c + FAR_c), where HR_c is the number of hits made with confidence level *c* divided by the number of targets and FAR_c is the number of false alarms made with confidence level *c* divided by the number of lures. The use of hit and false

alarm *rates* is what makes the CAC plot independent of base rate. Because, in the real world, the base rate of target-present lineups is unknown, this might be the best way to construct a CAC plot. Note that, unlike in the case of ROC analysis, these hit and false alarm rates are not cumulative over confidence. For example, if a 6-point rating scale is used (where ratings of 1 through 3 reflect different levels of confidence in a "no" decision and ratings of 4 through 6 reflect different levels of confidence in a "yes" decision), then HR_5 is equal to the number of hits made with a confidence rating of 5 divided by the number of stimulus-present trials, and FAR_5 is equal to the number of false alarms made with a confidence rating of 5 divided by the number of stimulus-absent trials.

It is worth emphasizing here that the CAC formula for positive suspect IDs made with confidence level c, namely $HR_c / (HR_c + FAR_c)$, could just as easily be expressed as a ratio, HR_c / FAR_c. This is the familiar diagnosticity ratio considered earlier in relation to Bayes' theorem. Although the DR is not a measure of discriminability (despite having been used for that purpose over many years), it is a perfectly valid measure for assessing the confidence-accuracy relationship. Indeed, studies that used calibration curves to shed light on the confidence-accuracy relationship have also sometimes reported the DR for each level of confidence. In other words, they also reported CAC data in ratio form (e.g., see Table 9 of Brewer & Wells, 2006), though without using it to gauge the confidence-accuracy relationship. Yet the information contained in this ratio is precisely the information one needs to determine if confidence in a suspect ID provides information about the accuracy of that ID. However, we much prefer the CAC formula because, in our experience, people find a proportion much easier to understand than an odds ratio.

Consider next what signal detection theory naturally predicts about the confidence-accuracy relationship, which is illustrated in Figure 11. For high-confidence "yes" decisions (confidence rating = 6), HR_6 = .31 and FAR_6 = .01 such that high-confidence accuracy is equal to .31 / (.31 + .01) = .98. In other words, decisions made with high confidence are correct 98% of

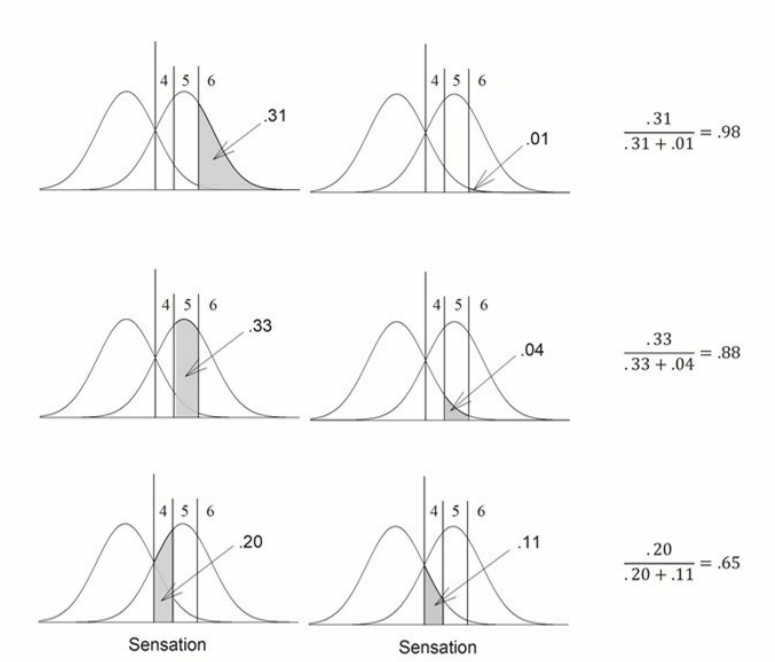

Figure 11. The same model as shown in Figure 10 except focusing on "Yes" decisions (confidence ratings of 4, 5 or 6), illustrating how signal detection theory predicts a strong confidence-accuracy relationship.

the time. For medium-confidence "yes" decisions (confidence rating = 5), HR_5 = .33 and FAR_5 = .04 such that medium-confidence accuracy is equal to .33 / (.33 + .04) = .88. For low-confidence "yes" decisions (confidence rating = 4), HR_4 = .20 and FAR_4 = .11 such that low-confidence accuracy is equal to .20 / (.20 + .11) = .65. Although not illustrated here, for parallel reasons, the model predicts a similar confidence-accuracy relationship for "no" decisions.

The same basic prediction applies to the predicted confidence-accuracy relationship for lineups except that the prediction is limited to positive IDs because no specific face is associated with No IDs. The key point is that, using signal detection theory as a guide, the best way to assess the confidence-accuracy relationship is to simply plot suspect ID accuracy (i.e., proportion

correct) as a function of confidence (Mickes, 2015). As noted earlier, this is precisely the information that is of interest to judges and jurors. Given that a suspect ID was made with a particular level of confidence, how likely is it that the ID is accurate? A CAC plot answers that question. Stated differently, a CAC plot (unlike an ROC plot) depicts what we think of as "reliability."

Recently, Wixted and Wells (2017) reviewed many laboratory studies and created their corresponding CAC plots. Figure 12A reproduces their summary figure based on 15 simulated crime studies. Obviously, high-confidence implies very high accuracy and low-confidence

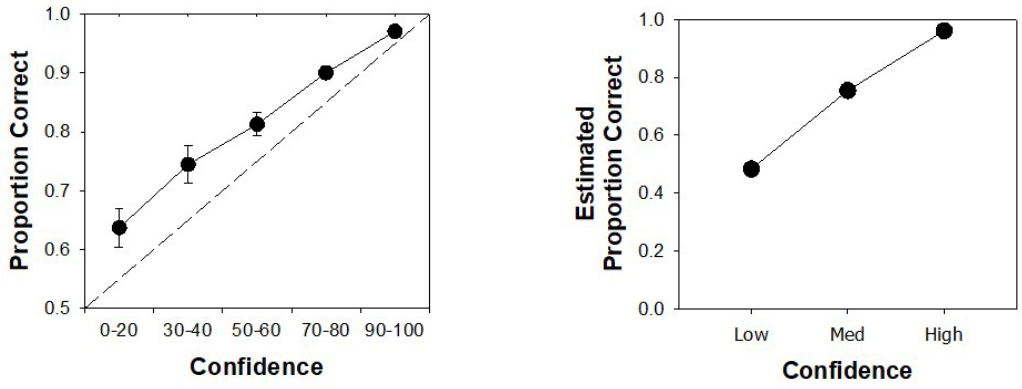

Figure 12. Left panel: CAC plot showing suspect ID accuracy (proportion correct) averaged across 15 studies with comparable scaling on the confidence (x) axis (Wixted & Wells, 2017). Right panel: Estimated suspect ID accuracy (proportion correct) as a function of confidence for the data from the Houston Police Department field study (Wixted, Mickes, Clark, Dunn, & Wells, 2016).

implies much lower accuracy. Wixted et al. (2016) reported similar results from the Houston Police Department field study, shown here in Figure 12B. The latter study is particularly important because it provides evidence that in actual practice (with real eyewitnesses), memory is reliable on an initial uncontaminated test using proper testing procedures.

These data suggest that, on the initial test, eyewitness memory is highly reliable in the sense that low confidence implies low accuracy (often not much better than chance) whereas high confidence implies high accuracy (often 95% correct or better). It is only later, after

memory has been contaminated, that confidence may no longer predict accuracy, and high-confidence errors may be common.

Some researchers are willing to accept the possibility that, in principle, when first tested under perfect ("pristine") testing conditions, eyewitness memory can be reliable. However, these same researchers worry that pristine testing conditions almost never occur in the real world (Berkowitz & Frenda, 2018; Wade, Nash, & Lindsay, 2018). For example, instead of using a fair lineup, perhaps the police often use an unfair lineup. If so, then it could be argued that although an initial test of eyewitness memory is theoretically reliable, in actual practice, it is unreliable, just as many have contended all along. However, the evidence weighs against this perspective. To be sure, the police should always use pristine testing procedures because there is no compelling reason not to (National Research Council, 2014; Wells et al., 2020). Indeed, c*ertain* non-pristine testing conditions clearly do compromise the information-value of even a high-confidence initial identification, most notably, the use of an unfair lineup in which the suspect, innocent or guilty, best matches the memory of the perpetrator (Wixted & Wells, 2017). However, as described next, the evidence to date suggests that *those* non-pristine conditions tend to be rare and that the non-pristine conditions that more commonly prevail may not seriously degrade the reliability of eyewitness memory when it is first tested.

Garrett (2011) analyzed trial materials for 161 DNA exonerees who had been misidentified by one or more eyewitnesses in a court of law and found that for every case in which initial eyewitness confidence could be determined (91 of 161 cases examined), the eyewitness appropriately expressed low confidence. Indeed, some eyewitnesses did not even identify the suspect on the first test (e.g., they rejected the lineup). Thus, despite the non-pristine testing conditions that Garrett (2011) also painstakingly documented for most of the DNA

exoneration cases, the initial eyewitness test result still came back as *inconclusive*. Conspicuously missing were initial misidentifications made with high confidence, which fits with other research showing that, even for real eyewitnesses, initial identifications from a lineup made with high confidence tend to be highly accurate (Wixted et al., 2016).

The importance of these facts cannot be overstated. For all other types of forensic evidence, an inconclusive test result would have been the end of it, but for eyewitness memory, the testing was repeated (often with confirming feedback) until the initially inconclusive low-confidence ID was transformed into a seemingly much more conclusive high-confidence ID by the time of the trial (Wells & Bradfield, 1998). To appreciate just how misguided this practice is, consider an analogy involving DNA evidence. Imagine that a forensic DNA sample yielded an inconclusive test result (e.g., too few alleles were recovered from the murder weapon to be of much use). Instead of accepting that outcome, imagine that law enforcement responded by (1) asking the suspect to handle the murder weapon, thereby depositing copious amounts of his DNA on it, (2) re-testing the now-contaminated evidence and finding a conclusive match to the suspect, and then (3) using that conclusive DNA match to convince a jury to return a guilty verdict. If such practices often resulted in the wrongful conviction of innocent suspects, the problem would not be that DNA evidence is unreliable. Instead, the problem would be that police and prosecutors ignored the inconclusive test result from the uncontaminated evidence and then used evidence that they themselves contaminated to help convict an innocent person.

This story of how DNA evidence can be misused is fictitious, but it corresponds exactly to the real story of how eyewitness evidence is misused. As noted above, in the wrongful conviction cases that are usually blamed on the unreliability of eyewitness memory, the initial test result came back as *inconclusive*. Unfortunately, that was not the end of it. Instead, well-

intentioned investigators tested the memory of the eyewitness again and again (each time further contaminating it and increasing confidence) until, ultimately, in front of a jury, the contaminated memory evidence seemed conclusive because the eyewitness identification of the innocent defendant was made with high confidence.

What sense does it make to blame the fallibility of eyewitness memory for these wrongful convictions (Wixted, 2018)? If a problem is incorrectly diagnosed, as it has been in the case of eyewitness memory for many years, the proposed solutions (e.g., "ignore eyewitness confidence") will be correspondingly off the mark. As it turns out, eyewitness memory was not especially problematic in *any* of the DNA exoneration cases for which the relevant information is available. The problem instead was that police and prosecutors did not accept the inconclusive result they obtained from the one and only uncontaminated test they conducted, as they would have had the forensic evidence involved DNA or fingerprints. Once appropriately diagnosed, the solution is simple: stop doing that. If a witness expresses low confidence in the initial ID, treat it as the weak evidence it is, no matter how confident they later become (no exceptions). Had that one simple rule been followed from the beginning, it is possible that none of the wrongful convictions ordinarily attributed to eyewitness misidentification would have occurred in the first place. In light of these considerations, a new verdict seems warranted: In addition to exonerating the innocent defendants who were wrongfully convicted, the time has come to exonerate eyewitness memory too (Wixted, 2018).

Estimator Variables and the Reliability of Eyewitness Identification

The discussion to this point has focused on "system variables," which are factors that are under the control of law enforcement (simultaneous vs. sequential lineups, fair vs. unfair lineups, etc.). As indicated earlier, estimator variables are factors that affect the accuracy of eyewitness

memory and that are outside the control of law enforcement. For example, consider a recent amicus brief filed by the American Psychological Association (APA) in the case of Commonwealth of Pennsylvania v. Walker (2014), which lists six estimator variables widely believed to affect the reliability of eyewitness identification:

1. *Passage of Time*.
2. *Witness Stress*.
3. *Exposure Duration*.
4. *Distance*.
5. *Weapon Focus*.
6. *Cross-Race Bias*.

On page 13, the upshot of the ostensible scientific consensus about these estimator variables is summarized as follows:

> The point is simply that eyewitness reliability— the linchpin of admissibility under this Court's precedent—is…determined by numerous factors identified by scientific research, many of which (the estimator variables) have nothing to do with the conduct of law enforcement. Eyewitness testimony can be unreliable even where there is no state-created suggestiveness.

This interpretation of how estimator variables affect the reliability of eyewitness identification probably comes as no surprise to the reader because it accords with textbook treatments of the issue. However, once again, this intuitively reasonable line of thinking is erroneous. As recently explained by Semmler, Dunn, Mickes, & Wixted (2018), the reason why becomes apparent by considering the theoretical interpretation of the relationship between suboptimal estimator variables and the reliability of eyewitness identification.

Theories of the Effect of Estimator Variables

Optimality Hypothesis. The main theory advanced to make sense of why all of these variables diminish the reliability of eyewitness identification is known as the "optimality hypothesis." The optimality hypothesis is not a statement of the reliability of eyewitness identifications, per se, but is instead a statement about the *correlation* between confidence and

accuracy under favorable vs. unfavorable information processing conditions. The proposal is that the confidence-accuracy correlation should vary directly with the optimality of those conditions (Deffenbacher, 1980). In other words, the correlation should be higher when (for example) exposure to the perpetrator is long, distance between the witness and perpetrator is short, and stress is low compared to when exposure to the perpetrator is short, distance between the witness and perpetrator is long, and stress is high. As Deffenbacher (2008) explained, under poor information processing conditions, "…not only will familiar faces be judged to be unfamiliar and unfamiliar faces be judged as familiar more frequently, but the same confidence rating is also more likely to be applied both to a judgment that a face seen before is indeed familiar and to a judgment that another face, never seen before, is also familiar" (p. 819). The optimality hypothesis therefore helps to explain the widespread belief that the usefulness of eyewitness confidence as an indicator of accuracy – including high confidence – will decrease as information processing conditions get worse.

Constant likelihood ratio signal detection model. When the estimator variable issue is conceptualized in terms of signal detection theory, it becomes clear that two distinct measures of accuracy have long been conflated, one of which (the irrelevant one) is indeed affected by suboptimal estimator variables and the other of which (the relevant one) is not (Semmler et al., 2018).

As noted earlier, in signal detection theory, confidence ratings are conceptualized in terms of different decision criteria. This is true of all of the specific models we considered (i.e., the Independent Observations model, the Ensemble model, and the Integration model). In essence, the likelihood ratio version of these models makes a prediction about how the confidence criteria shift on the memory-strength axis across conditions that affect *d′* (such as short vs. long exposure, same-vs.-cross race, etc.). Figure 13 shows a high *d′* condition and a low *d′* condition, with three confidence criteria (*c1*, *c2* and *c3*) for each condition. Note that the criteria are shifted (more spread out) when *d′* is low. In fact, each criterion is placed in such a way as to maintain a constant likelihood ratio across conditions. Consider for example, *c1*,

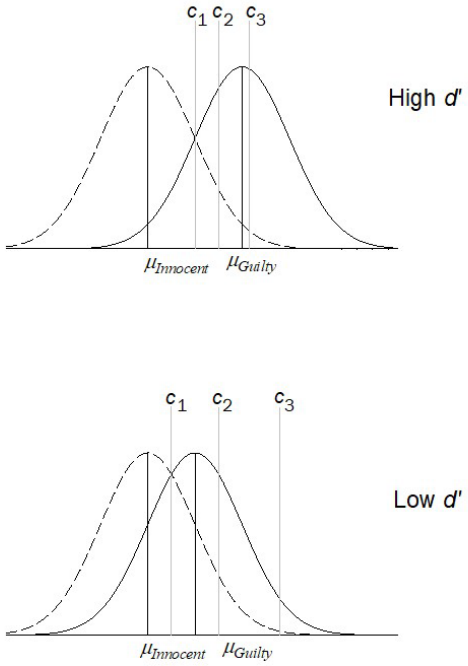

Figure 13. An illustration of the constant likelihood ratio signal detection model. For both the high *d′* condition and the low *d′* condition, the likelihood ratio of each confidence criterion remains unchanged. For example, the likelihood ratio for *c*1 is equal to 1 in both cases, and the likelihood ratio for *c*3 is equal to 10 in both cases.

which, in both conditions, is placed at the point where the height of the guilty suspect distribution is equal to the height of the innocent suspect distribution (likelihood ratio = 1). By

contrast, *c3* is placed at the point where the height of the guilty suspect distribution is ten times the height of the innocent suspect distribution (likelihood ratio = 10) in both conditions.

The models shown in Figure 13 illustrate the specific prediction made by constant likelihood ratio theories. Such theories, in one form or another, have long been a cornerstone in the basic recognition memory literature (Glanzer & Adams, 1990; Glanzer, Adams, Iverson & Kim, 1993; McClelland & Chappel, 1998; Osth, Dennis & Heathcote, 2017; Shiffrin & Steyvers, 1997; Stretch & Wixted, 1998; Wixted & Gaitan, 2002). The critical prediction is that the CAC curve should remain essentially unchanged as *d′* decreases (e.g., due to poor estimator variables). Thus, for example, although fewer high-confidence IDs will be made in the low-*d′* condition relative to the high-*d′* condition (e.g., in Figure 13, very few innocent and guilty suspects generate signals that exceed c3 in the low-*d′* condition), those that are made will nevertheless still be about ten times as likely to be correct than incorrect. Thus, high-confidence accuracy should remain essentially unchanged.

Semmler et al. (2018) recently tested this prediction by reanalyzing data from an earlier study in which distance was manipulated (i.e., witnesses were either close to or far away from the target while witnessing the event) and memory was measured using a lineup. The ROC data on the left (Figure 14A) clearly show the expected effect on discriminability, which short distance between the perpetrator and witness leading to much better performance than long distance. However, the CAC curves on the right (Figure 14B) show that the probability that a suspect ID was correct given that it occurred was unchanged. Moreover, high-confidence accuracy was very high in both conditions. Critically, this same pattern has been observed by multiple independent labs for a wide variety of estimator variables,

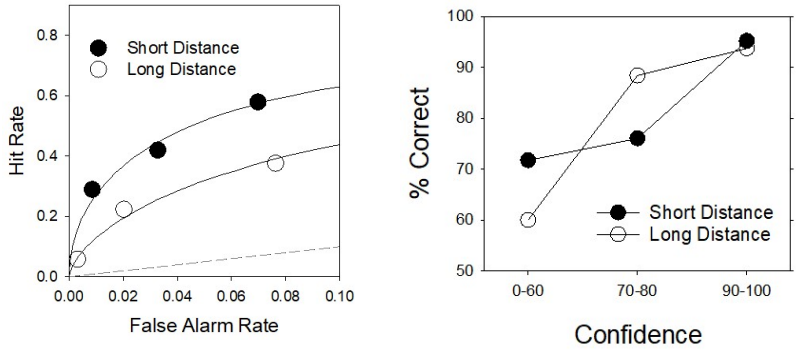

Figure 14. Left panel: ROC data for short and long distances in the delayed condition of Lindsay et al. (2008). Right panel: Suspect ID accuracy as a function of confidence when the distance between the witness and perpetrator was short vs. when it was long in the delayed condition of Lindsay et al. (2008). These re-analyses of the data reported by Lindsay et al. (2008) were conducted by Semmler et al. (2018).

including cross-race IDs (Dobolyi & Dodson, 2013; Nguyen, Pezdek, & Wixted, 2017), weapon focus (Carlson, Dias, Weatherford, & Carlson, 2017), retention interval (Palmer, Brewer, Weber, & Nagesh, 2013; Wixted, Read, & Lindsay, 2016) and exposure duration (Palmer et al., 2013). Thus, the prior verdict on estimator variables (namely, that suboptimal encoding conditions compromise the confidence-accuracy relationship) was incorrect.

To be sure, there is undoubtedly some low level of discriminability where the confidence-accuracy relationship will break down. To take an extreme, if conditions are so poor that $d' = 0$, then suspect ID accuracy will be at chance for all levels of confidence. If humans were *perfect* likelihood ratio computers, high-confidence IDs would never be made under those conditions, but humans are clearly not perfect likelihood ratio computers (Stretch & Wixted, 1998). Instead, their performance is reasonably well approximated by a constant likelihood ratio model. The key point is that over the range of d' values investigated in the laboratory, once initial confidence is known, the estimator variables provide little additional information.

The lab-based data help to make sense of what would otherwise be an almost inconceivable outcome, namely the high degree of reliability exhibited by the real eyewitnesses studied by Wixted et al. (2016). That study was conducted in the Robbery Division of the

Houston Police Department. A survey of cases from that division from the previous year reported by W. Wells, Campbell, Li and Swindle (2016) indicated that 61.6% were cross-race cases, 73.5% involved the presence of a weapon, and the average delay between the offense and the identification procedure was over a month (median = 2.5 weeks). Moreover, it seems reasonable to suppose that witness stress was typically high. In other words, the estimator variables (high stress, cross-race, long retention interval) were such that one might reasonably assume that the reliability of eyewitness identifications in this study would be very poor. Contrary to that assumption, estimated suspect ID accuracy was very similar to what has been observed in lab studies, and estimated high-confidence accuracy was very high (as illustrated earlier in Figure 12B).

Conclusion

If there is one message that we have tried to convey here is that applied science can go astray without guidance from the theoretical models that have been painstakingly worked out by basic scientists. In the absence of considerations from basic science in general—and signal detection theory in particular—the message sent by applied psychologists was that the police should use sequential lineups, that eyewitness memory is unreliable even when initial confidence is high, and that it becomes even more unreliable when suboptimal estimator variables prevail. But a completely different message emerges once these issues are considered in the light of signal detection theory.

As it turns out, ROC analysis reveals that simultaneous lineups are diagnostically superior to sequential lineups, eyewitness identification is highly reliable on an initial test using a fair lineup, and reliability remains high even when the estimator variables are suboptimal. Further, from the work of Elizabeth Loftus and others, we have long known that memory is

malleable. Indeed, because the very act of testing memory contaminates it, all subsequent identification tests beyond the initial test are less informative, especially the one that occurs at the time of trial (which is when highly publicized high-confidence misidentifications often occur). Critically, the fact that the legal system tends to ignore the reliable-but-inconclusive first ID and to focus on the unreliable-but-conclusive later ID is not the fault of the eyewitness. Like other kinds of forensic evidence (e.g., DNA), eyewitness memory is highly reliable when it is not contaminated. Thus, it is not eyewitnesses who are unreliable. Instead, it is the legal system itself that is unreliable because it chooses to rely on contaminated forensic evidence by asking witnesses to identify the perpetrator at the time of the trial and to express confidence in that ID. This is precisely how the many DNA exonerees ended up in prison in the first place (none of whom, so far as we know, was misidentified with high-confidence at the time of the initial ID).

The solution to that tragic problem is painfully simple: stop doing that and focus on the initial ID instead. Although this solution is as easy to understand as it is to implement, it would not have been appreciated without the guidance provided by signal detection theory. Indeed, the recent developments in the field of eyewitness memory provide a clear illustration of the fact that basic science and applied science have become far too detached. As a first step in bringing them closer together, we suggest that graduate training programs should not award the Ph.D. unless the student exhibits some degree of proficiency in signal detection theory and ROC analysis. Such proficiency is arguably more important than proficiency in null hypothesis significance testing, yet every graduate student is trained in the latter and few are trained in the former. It might be time to reconsider our graduate training priorities.

References

Albright, T. D. & Rakoff, J. S. (2020). The impact of the National Academy of Sciences report on eyewitness identification. *Judicature, 104*, 21-29.

Amendola, K. L. & Wixted, J. T. (2015a). Comparing the diagnostic accuracy of suspect identifications made by actual eyewitnesses from simultaneous and sequential lineups in a randomized field trial. *Journal of Experimental Criminology, 11*, 263-284.

Amendola, K. L. & Wixted, J. T. (2015b). No possibility of a selection bias, but direct evidence of a simultaneous superiority effect: a reply to Wells et al. *Journal of Experimental Criminology, 11*, 291-294.

Amendola, K. L. & Wixted, J. T. (2017). The role of site variance in the American Judicature Society field study comparing simultaneous and sequential lineups. *Journal of Quantitative Criminology, 33*, 1-19.

*Andersen, S. M., Carlson, C. A., Carlson, M. & Gronlund, S. D. (2014). Individual Differences Predict Eyewitness Identification Performance. *Personality and Individual Differences, 60*, 36-40.

APA brief for Commonwealth v. Walker (2014). Retrieved from: http://www.apa.org/about/offices/ogc/amicus/walker.aspx

Berkowitz, S. & Frenda, S. Rethinking the confident eyewitness: A reply to Wixted, Mickes, and Fisher. *Perspectives on Psychological Science* **13** (2018) 336-338.

Brewer, N., Keast, A., & Rishworth, A. (2002). The confidenceaccuracy relationship in eyewitness identification: The effects of reflection and disconfirmation on correlation and calibration. *Journal of Experimental Psychology: Applied, 8*, 44–56.

Brewer, N., & Wells, G. L. (2006). The confidence-accuracy relation in eyewitness identification: Effects of lineup instructions, foil similarity, and target-absent base rates. *Journal of Experimental Psychology: Applied, 12*, 11–30

*Carlson, C. A. & Carlson, M. A. (2014). An evaluation of perpetrator distinctiveness, weapon presence, and lineup presentation using ROC analysis. *Journal of Applied Research in Memory and Cognition, 3, 45-53. doi:10.1016/j.jarmac.2014.03.004*

Carlson, C. A., Dias, J. L., Weatherford, D. R., & Carlson, M. A. (2017). An investigation of the weapon focus effect and the confidence–accuracy relationship for eyewitness identification. *Journal of Applied Research in Memory and Cognition, 6,* 82-92.

Ceci, S. J., & Bruck, M. (1993). Suggestibility of the child witness: A historical review and synthesis. *Psychological Bulletin*, *113*, 403–439.

Ceci, S. J., Loftus, E. F., Leichtman, M. D., & Bruck, M. (1994). The possible role of source misattributions in the creation of false beliefs among preschoolers. *International Journal of Clinical and Experimental Hypnosis*, *42*, 304–320.

Clark, S. E. (2003). A memory and decision model for eyewitness identification. *Applied Cognitive Psychology, 17*, 629–654.

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

Clark, S. E., Erickson, M. A., & Breneman, J. (2011). Probative value of absolute and relative judgments in eyewitness identification. *Law and Human Behavior, 35*, 364–380.

Colloff, M. F., Wade, K. A., & Strange, D. (2016). Unfair lineups make witnesses more likely to confuse innocent and guilty suspects. *Psychological Science, 27,* 1227–1239.

Colloff, M. F., Wade, K. A., Strange, D., & Wixted, J. T. (2018). Filler Siphoning Cannot

Explain The Fair Lineup Advantage: Reply to Smith, Wells, Smalarz, and Lampinen (2018). *Psychological Science, 29*, 1552-1557.

Colloff, M. F. & Wixted, J. T. (2019). Why are Lineups Better than Showups? A Test of the Filler Siphoning and Enhanced Discriminability Accounts. *Journal of Experimental Psychology: Applied*. doi: 10.1037/xap0000218

Davis, D. & Loftus, E. F. (2018). Eyewitness Science in the 21st Century: What Do We Know and Where Do We Go from Here? In J. T. Wixted (Editor) and E. A. Phelps & L. Davachi (Volume Editors) *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience: Learning and Memory* (4th edition, Volume 1). Wiley.

Deffenbacher, K. A. (1980). Eyewitness accuracy and confidence: Can we infer anything about their relationship? *Law and Human Behavior, 4*, 243–260.

Deffenbacher, K. A. (2008). Estimating the impact of estimator variables on eyewitness identification: A fruitful marriage of practical problem solving and psychological theorizing? *Applied Cognitive Psychology, 22,* 815–826. http://dx.doi.org/10.1002/acp.1485

Dobolyi, D. G. & Dodson, C. S. (2013). Eyewitness confidence in simultaneous and sequential lineups: a criterion shift account for sequential mistaken identification overconfidence. *Journal of Experimental Psychology: Applied, 19*, 345–357.

Duncan, M. (2006). *A signal detection model of compound decision tasks.* (Tech Note DRDC TR 2006-256. Toronto, Defence Research and Development Canada.

Egan, J. P. (1958). *Recognition memory and the operating characteristic*. (Tech Note AFCRC-TN-58-51). Bloomington, IN: Indiana University, Hearing and Communication Laboratory.

Fechner, G. T. (1966). *Elements of psychophysics* (Vol. I) (E.G. Boring & D.H. Howes, Eds.; H.E. Adler, Trans.). New York: Holt. Rinehart & Winston. (Original work published 1860).

Fife, D., Perry, C., & Gronlund, S. D. (2014). Revisiting absolute and relative judgments in the WITNESS model. *Psychonomic Bulletin & Review, 21*, 479-487.

Fitzgerald, R. J., Price, H. L., Oriet, C., & Charman, S. D. (2013). The effect of suspect-filler similarity on eyewitness identification decisions: A meta-analysis. *Psychology, Public Policy, and Law, 19*, 151–164.

Garrett, B. (2011). *Convicting the Innocent: Where Criminal Prosecutions Go Wrong*. Cambridge, MA: Harvard University Press.

Gibson, E. J. (1969). *Principles of perceptual learning and development*. New York, NY: Appleton-Century-Crofts.

Glanzer, M., & Adams, J. K. (1990). The mirror effect in recognition memory: Data and theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*, 5–16.

Glanzer, M., Adams, J. K., Iverson, G. J., & Kim, K. (1993). The regularities of recognition memory. *Psychological Review, 100*, 546–567.

*Goodsell, C. A. (unpublished). Effects of eyewitness memory encoding strength on sequential and simultaneous lineup identifications.

Goodsell, C. A., Gronlund, S. D., & Carlson, C. A. (2010). Exploring the sequential lineup advantage using WITNESS. *Law and Human Behavior, 34*, 445–459.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics* (A reprint, with corrections of the original 1966 ed.). Huntington, NY: Robert E. Krieger Publishing Co.

Gronlund, S. D. & Benjamin, A. S. (2018). The new science of eyewitness memory. *Psychology*

of Learning and Motivation-Advances in Research and Theory, 69, 241-284.

*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.

Gronlund, S. D., Wixted, J. T. & Mickes, L. (2014). Evaluating eyewitness identification procedures using ROC analysis. *Current Directions in Psychological Science, 23*, 3-10.

Horry, R., Brewer, N., Weber, N. & Palmer, M. (2015). The effects of allowing a second sequential lineup lap on choosing and probative value. *Psychology, Public Policy, and Law 21*, 121-133.

Innocence Project (2019). Understand the causes: the causes of wrongful conviction. New York: Innocence Project. https://www.innocenceproject.org/eyewitness-identification-reform/. Accessed August 31, 2019.

Juslin, P., Olsson, N., & Winman, A. (1996). Calibration and diagnosticity of confidence in eyewitness identification: Comments on what can be inferred from the low confidence-accuracy correlation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*, 1304-1316.

Lindsay R. C., & Wells, G. L. (1985). Improving eyewitness identifications from lineups: Simultaneous versus sequential lineup presentation. *Journal of Applied Psychology, 70,* 556-564.

Loftus, E. F. (2003). Make-believe memories. *American Psychologist, 58*, 867-873.

Loftus, E., & Ketcham, K. *The myth of repressed memory: False memories and allegations of sexual abuse*. (1994) New York: St. Martin's Press.

Loftus, E. F., Miller, D. G., & Burns, H. J. (1978). Semantic integration of verbal information

into a visual memory. *Journal of Experimental Psychology: Human Learning and Memory, 4*, 19–31.

Loftus, E. F. & Palmer, J. C. (1974) Reconstruction of automobile destruction: An example of the interaction between language and memory. J*ournal of Verbal Learning and Verbal Behavior, 13*, 585 -589.

Loftus, E. F. & Pickrell, J. E. (1995). The formation of false memories. *Psychiatric Annals, 25*, 720–725.

Ma, Correll, & Wittenbrink (2015). The Chicago Face Database: A Free Stimulus Set of Faces and Norming Data. *Behavior Research Methods, 47,* 1122-1135.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide (2nd ed.)*. New York, NY: Cambridge University Press.

McClelland, J. L., & Chappell, M. (1998). Familiarity breeds differentiation: A subjective-likelihood approach to the effects of experience in recognition memory. *Psychological Review, 105*, 724–760.

*Meisters, J., Diedenhofen, B., & Musch, J. (2018). Eyewitness identification in simultaneous and sequential lineups: An investigation of position effects using receiver operating characteristics. *Memory, 26*, 1297–1309.

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 & Cognition , 4*, 93–102.

*Mickes, L., Flowe, H. D. & Wixted, J. T. (2012). Receiver operating characteristic analysis of eyewitness memory: Comparing the diagnostic accuracy of simultaneous and sequential lineups. *Journal of Experimental Psychology: Applied, 18*, 361-376.

Mickes, L., Seale-Carlisle, T., Wetmore, S., Gronlund, S., Clark, S., Carlson, C. A., Goodsell, C., Weatherford, D. & Wixted, J. T. (2017). ROCs in eyewitness identification: Instructions versus confidence ratings. *Applied Cognitive Psychology*. http://dx.doi.org/10.1002/acp.3344

Moreland, M. B. & Clark, S. E. (2019). Absolute and relative decision processes in eyewitness identification. *Applied Cognitive Psychology*, https://doi.org/10.1002/acp.3602.

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

Neuschatz, J. S., Wetmore, S. A., Key, K., Cash, D., Gronlund, S. D., & Goodsell, C. A. (2016). A Comprehensive evaluation of showups. In M. Miller & B. Bornstein (Eds.), *Advances in Psychology and Law* (vol. 1, pp. 43–69). Springer.Osth, A.F., Dennis, S., & Heathcote, A. (2017). Likelihood ratio sequential sampling models of recognition memory. *Cognitive Psychology, 92*, 101-126.

Nguyen, T. B., Pezdek, K. & Wixted, J. T. (2017). Evidence for a Confidence-Accuracy Relationship in Memory for Same- and Cross-Race Faces. *Quarterly Journal of Experimental Psychology, 70*, 2518-2534.

Palmer, M. A. & Brewer, N. (2012). Sequential lineup presentation promotes less-biased criterion setting but does not improve discriminability. *Law and Human Behavior, 36,* 247-255.

Palmer, M. A., Brewer, N. & Horry, R. (2013). Understanding gender bias in face recognition: Effects of divided attention at encoding. *Acta Psychologica 142*, 362-369.

Palmer, M. A., Brewer, N., & Weber, N. (2010). Postidentification feedback affects subsequent eyewitness identification performance. *Journal of Experimental Psychology: Applied, 16*, 387-398.

Palmer, M., Brewer, N., Weber, N., & Nagesh, A. (2013). The confidence–accuracy relationship for eyewitness identification decisions: Effects of exposure duration, retention interval, and divided attention. *Journal of Experimental Psychology: Applied, 19*, 55–71.

Penrod, S., & Cutler, B. (1995). Witness confidence and witness accuracy: Assessing their forensic relation. *Psychology, Public Policy, and Law, 1*, 817–845.

Police Executive Research Forum (2013). A National Survey of Eyewitness Identification Procedures in Law Enforcement Agencies. Retrieved March 29, 2016, from http://www.policeforum.org/

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J., & Müller, M. (2011). pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics, 12*, 77.

*Rotello, C., Guggenmos, & Isbell (unpublished).

Seale-Carlisle, T. M., & Mickes, L. (2016). U.S. line-ups outperform U. K. line-ups. *Royal Society Open Science, 3*, 160300.

*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.

Semmler, C., Dunn, J., Mickes, L. & Wixted, J. T. (2018). The Role of Estimator Variables in Eyewitness Identification. *Journal of Experimental Psychology: Applied, 24*, 400-415.

Shiffrin, R.M. & Steyvers, M. (1997). A model for recognition memory: REM: Retrieving Effectively from Memory. *Psychonomic Bulletin & Review, 4*, 145-166.

Smith, A. M., Lampinen, J. M., Wells, G. L., Smalarz, L., &. Mackovichova, S. (2018). Deviation from Perfect Performance Measures the Diagnostic Utility of Eyewitness Lineups but Partial Area Under the ROC Curve Does Not. J*ournal of Applied Research in Memory and Cognition, 8*, 50–59.

Smith, A. M., Wells, G. L., Lindsay, R. C. L., & Penrod, S. D. (2017). Fair lineups are better than biased lineups and showups, but not because they increase underlying discriminability. *Law and Human Behavior, 41*, 127-145.

Smith, A. M., Wells, G. L., Smalarz, L., & 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*, 1548–1551.

Steblay, N. K., Dysart, J. E., & Wells, G. L. (2011). Seventy-two tests of the sequential lineup superiority effect: A meta-analysis and policy discussion. *Psychology, Public Policy, and Law, 17,* 99-139.

Stretch, V., & Wixted, J. T. (1998). Decision rules for recognition memory confidence judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*, 1397–1410.

Swets, J. A., Tanner, W. P., & Birdsall, T. G. (1955). The evidence for a decision making theory of visual detection. Technical Report No. 40, University of Michigan, Electronic Defense Group.

Tanner, W. P. & Swets. J. A., & Green, D. M. (1956). Some general properties of the hearing mechanism. Technical Report No. 30, University of Michigan, Electronic Defense Group.

*Terrell, J. T., Baggett, A. R., Dasse, M. N., & Malavanti, K. F. (2017). A hybridization of simultaneous and sequential lineups reveals diagnostic features of both traditional procedures. *Applied Psychology in Criminal Justice, 13*, 96-109.

Tversky, A. (1977). Features of similarity. *Psychological Review, 84*, 327–352.

Wade, K., Nash, R. & Lindsay, D. S. (2018). Reasons to doubt the reliability of eyewitness memory: Commentary on Wixted, Mickes, and Fisher (2018). *Perspectives on Psychological Science* **13** (2018) 339-32.

Weber, N., & Brewer, N. (2004). Confidence-accuracy calibration in absolute and relative face recognition judgments. *Journal of Experimental Psychology: Applied, 10*, 156–172.

Wells, G. L. (1978). Applied eyewitness testimony research: System variables and estimator variables. *Journal of Personality and Social Psychology, 36,* 1546–1557.

Wells, G. L. (1984). The psychology of lineup identifications. *Journal of Applied Social Psychology, 14*, 89–103.

Wells, G. L., & Bradfield, A. L. (1998). "Good, you identified the suspect": Feedback to eyewitnesses distorts their reports of the witnessing experience. *Journal of Applied Psychology*, *83*, 360-376.

Wells, G.L., Dysart, J.E. & Steblay, N.K. (2015). The flaw in Amendola and Wixted's conclusion on simultaneous versus sequential lineups. *Journal of Experimental Criminology, 11*, 285–289.

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*, 3-36.

Wells, G. L., & Murray, D. M. (1984). Eyewitness confidence. In G. L. Wells & E. F. Loftus (Eds.), *Eyewitness testimony: Psychological perspectives* (pp. 155–170). New York, NY: Cambridge University Press.

Wells, W., Campbell, B., Li, Y. & Swindle, S. (2016). The characteristics and results of eyewitness identification procedures conducted during robbery investigations in Houston, TX. *Policing: An International Journal of Police Strategies & Management, 39*, 601 – 619.

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.

Wetmore, S. A., McAdoo, R. M., Gronlund, S. D., & Neuschatz, J. S., (2017). The impact of fillers on lineup performance. *Cognitive Research: Principles and Implications, 2*, 48.

Wetmore, S. A., Neuschatz, J. S., Gronlund, S. D, Goodsell, C. A., Wooten, A., & Carlson, C. A. (2015a). Effect of retention interval on showup and lineup performance. *Journal of Applied Research in Memory and Cognition, 4*, 8–14.

*Willing, S., Diedenhofen, B., & Musch, J. (under review). Presenting a similar foil prior to the suspect reduces the identifiability of the perpetrator in sequential lineups: A ROC-based

analysis.

Wixted, J. T. (2018). Time to exonerate eyewitness memory. *Forensic Science International*, e13-e15. doi: 10.1016/j.forsciint.2018.08.018

Wixted, J. T. (2019). The forgotten history of signal detection theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. doi: 10.1037/xlm0000732

Wixted, J. T. & Gaitan, S. (2002). Cognitive theories as reinforcement history surrogates: The case of likelihood ratio models of human recognition memory. *Animal Learning & Behavior, 30*, 289-305.

Wixted, J. T. & Mickes, L. (2014). A signal-detection-based diagnostic-feature-detection model of eyewitness identification. *Psychological Review, 121,* 262-276.

Wixted, J. T. & Mickes, L. (2018). Theoretical vs. empirical discriminability: the application of ROC methods to eyewitness identification. *Cognitive Research: Principles and Implications* 3:9.

Wixted, J. T., Mickes, L., Wetmore, S. A., Gronlund, S. D. & Neuschatz, J. S. (2017). ROC Analysis in Theory and Practice. *Journal of Applied Research in Memory and Cognition,* 6, 343-351.

Wixted, J. T., Mickes, L., Dunn, J., Clark, S. E., & Wells, W. (2016). Relationship between confidence and accuracy for eyewitness identifications made from simultaneous and sequential police lineups. *Proceedings of the National Academy of Sciences, 113,* 304-309.

Wixted, J. T., Mickes, L. & Fisher, R. P. (2018). Rethinking the Reliability of Eyewitness Memory. *Perspectives on Psychological Science, 13*, 324-335.

Wixted, J. T., Read, J. D. & Lindsay, D. S. (2016). The effect of retention interval on the eyewitness identification confidence-accuracy relationship. *Journal of Applied Research in Memory and Cognition, 5*, 192-203.

Wixted, J. T. & Wells, G. L. (2017). The Relationship between Eyewitness Confidence and Identification Accuracy: A New Synthesis. *Psychological Science in the Public Interest, 18*, 10-65.

Wixted, J. T., Vul, E., Mickes, L. & Wilson, B. W. (2018). Models of lineup memory. *Cognitive Psychology, 105*, 81-114.