Eyewitness Identification

Laura Mickes
Royal Holloway, University of London, United Kingdom
Department of Psychology
laura.mickes@rhul.ac.uk

Scott D. Gronlund
University of Oklahoma, United States
Department of Psychology
gronlund@ou.edu
Abstract

We present results that challenge the notion that eyewitness memory is notoriously inaccurate. The core of this challenge arises from viewing research from the joint perspectives of discriminability (the ability to distinguish innocent from guilty suspects) and reliability (the likelihood the person identified is actually the perpetrator). We also introduce a new organizational framework that classifies variables according to the different stages of memory as they map onto the different stages of the crime/criminal justice process. The discriminability/reliability perspective and the new organizational framework have major implications for how eyewitness identification evidence should be utilized by the criminal justice system.

Acknowledgements

This work was partially supported by the Economic and Social Research Council [ES/L012642/1] to Laura Mickes.

We thank Kayolan Ganev for his valuable assistance.
1 Introduction

Eyewitnesses have a bad rap for being untrustworthy. Although eyewitnesses have difficulty withstanding suggestions and feedback, their memories also have been faulted due to an inability to use confidence levels to reflect the likely accuracy of their identification. After all, faulty eyewitness identifications have played a role in 70% of the wrongful imprisonment of the innocent (Innocence Project, 2015). A glance at the eyewitness literature, the media coverage, and what is written in textbooks, makes it seem as if the memory of an eyewitness is functionally useless. Take, for example, the statement written in a recent news article covering a crime in which there were multiple eyewitnesses: “While eyewitness accounts are often fickle – and even shockingly inaccurate – the immediacy of the response in this case helped police act quickly. The video was certainly key, given the unreliability of human memory” (The Christian Science Monitor, 2016). Is the notorious reputation merited? Can eyewitness memory never be trusted? This chapter presents results that challenge that notion.

1.1 System vs. Estimator Variables

Researchers have dissected the variables that affect eyewitness accuracy and studied each constituent part. Those parts often have then been grouped into whether they can be used to make decisions to improve the system (e.g., decisions made regarding how to conduct a lineup, or what instructions to read to an eyewitness) or whether they involve variables that affect the accuracy of an identification (e.g., a cross-race identification, length of the retention interval). These variables typically have been categorized as system and estimator variables (Wells, 1978), respectively. A system variable is a variable that is under the control of the legal system and an estimator
variable is not. That distinction, however, only applies to policymakers who need to make decisions about how eyewitness evidence should be gathered by the police, because how a lineup is conducted can be specified, but whether the eyewitness and the perpetrator are the same race cannot. However, there is no such system/estimator distinction when it comes to the court of law. That is, for judges and jurors who need to determine the likely culpability of a defendant based on the identification of an eyewitness, there is no control over any variable whatsoever. Take, for example, lineup presentation. Whether to present an eyewitness a simultaneous- or sequentially-presented procedure is under the control of the police, and thus is a system estimator at the investigative stage. But when that case moves to the court of law, whether a simultaneous or sequential lineup was presented is a variable that is not under the control of the judge and jurors.

1.2 A new conceptualization of variables that affect memory

Figure 1 depicts a new way to conceptualize the variables that affect eyewitness memory, and how, over the course of a crime and its subsequent investigation, the nature of the variables changes. Variables are categorized according to its occurrence in a stage in memory and the stage in the crime and subsequent criminal proceedings. The events of the crime are encoded, stored (maintained in memory) during the time between the crime and the reporting of the crime, and finally the memory is retrieved when reporting the crime and attempting to identify the perpetrator from a lineup.

To provide context, consider the following example in which a robbery is witnessed. If the perpetrator is of a different race than the witness, then that may affect later identification because of the cross race bias (Meissner & Brigham, 2001, a phenomenon thought to occur during Encoding). The time between witnessing the robbery and reporting the crime (Storage) would also affect memory because longer
retention intervals will decrease memory performance due to the effects of forgetting and interference and create opportunities for influences such as misinformation (e.g., Loftus, 2005) and suggestibility (e.g., Zaragoza & Lane, 1994; Lindsay & Johnson, 1989). When the witness attempts to identify the perpetrator from a lineup (Retrieval), the type of lineup, such as a simultaneous lineup, may make a difference in memory performance compared to if memory was tested on, say, a sequential lineup.

1.3 Two Types of Accuracy

The organizational structure in Figure 1 provides a roadmap for the types of analyses that are important to the stakeholders that deal with eyewitness evidence. Both sets of stakeholders are interested in accuracy, but in different types of accuracy, and therefore, results from different types of analyses are relevant. One type of accuracy, the type that is most often considered in laboratory list-learning memory experiments, is discriminability. In list-learning experiments, a list of items, called targets, are presented during the study phase and then, during the test phase, targets and items that were not presented during the study phase, called lures, are randomly intermixed and presented to a participant. The participant indicates whether each item was or was not on the list. Likewise, in a forensically-relevant experiment, during the study phase, a participant will typically view a video of a mock crime where the perpetrator is the target. Memory for the perpetrator is then tested using a lineup.

Discriminability is the ability to distinguish the target, or the guilty suspect (the individual who committed the crime), from a lure, or the innocent suspect (an individual who the police incorrectly suspect of committing the crime). Discriminability is typically measured by $d'$ (a parametric measure of discriminability, or sensitivity), percent correct,
or area under the Receiver Operating Characteristic (ROC) curve (a non-parametric measure of discriminability) (see Macmillan & Creelman, 2005). Approaches to measuring lineup discriminability are discussed in detail in section 2.2.

A different type of accuracy – one that should be considered when determining eyewitness reliability (i.e., the likelihood that the identified suspect is guilty) – applies to subsets of eyewitnesses with the same discriminability. To evaluate accuracy of this type, positive predictive value (for example, the proportion correct or a likelihood ratio at each level of confidence) is considered. This is typically measured with calibration analysis (Juslin, Olsson, & Winman, 1996), or a close relative, confidence-accuracy characteristic (CAC) analysis (Mickes, 2015). Calibration analysis, especially CAC analysis, provides the information that is most relevant to judges and jurors, who are charged with determining culpability, because it indicates the accuracy for a given level of confidence reported by the eyewitness. Approaches to measuring calibration are discussed in detail in section 2.3.

Decisions about variables that are part of the investigation process, that might increase discriminability, should be informed by results from ROC analysis (as depicted in Figure 1). This is the type of accuracy that policymakers must consider when making decisions regarding whether one identification procedure is superior to another. However, courtroom decisions about eyewitness reliability should be informed by results from calibration-type analyses. When it comes to advising legal professionals, or testifying in courts as expert witnesses, this distinction must be kept in mind.

We begin our chapter by expanding upon the assessment of discriminability and reliability. As mentioned, discriminability can be assessed using ROC analysis, and we will introduce this technique and discuss several arguments that have been made regarding its use. This will be followed by a discussion of ways to measure positive predictive value to
assess reliability. Next, we will turn to a review of several variables known to affect eyewitness memory, and will consider the impact of these variables on discriminability and reliability. These variables will be organized according to the memory phase upon which they have the biggest impact (encoding, storage, or retrieval). We will close with the implications of the changing views of eyewitness identification revealed by the assessment of discriminability using ROC analyses and the assessment of the reliability using calibration and CAC analyses.

2  Testing and Measuring Eyewitness Memory

Eyewitness identification provides evidence during the course of criminal cases. One way police investigators glean information from an eyewitness is to administer a type of identification procedure, which is essentially a recognition memory test of the eyewitness’s memory for the perpetrator. Identification procedures include lineups (photos or videos), live lineups, showups, field views, and mugbooks (National Research Council, 2014). The variables that affect eyewitness accuracy are often measured in the lab using lineups or showups, and that is what we focus on in this chapter.

2.1  Lineups

A standard police lineup contains one suspect (who is either guilty or innocent) and multiple fillers (people who are known to be innocent). Figure 2 shows a schematic of a 6-person simultaneous photo lineup. The images of each lineup member (photos or videos) are either simultaneously or sequentially presented, depending on the policy adopted by a particular country or by a particular jurisdiction within a country.

<Figure 2 near here>
There are three possible choices that eyewitnesses can make when faced with a lineup. They can identify 1) the suspect, 2) one of the fillers, or 3) no one. All possible outcomes are in Table 1. If the target is present in the lineup, and the eyewitness identifies them, then that is a correct identification (also known as a hit). If the target is not present in the lineup, and the eyewitness identifies the innocent suspect, then that is a false identification (also known as a false alarm). These decisions are in red bolded font because these are the decisions that matter, and thus the focus is typically placed on the identification of the guilty and innocent suspects. That is because fillers are known to be innocent, and although picking one is an error, there is no risk of the filler being wrongfully investigated or convicted. If the eyewitness does not identify a lineup member, and if the target is in the lineup, then that is a miss; and if the target is not present in the lineup, then that is a correct rejection. In real police investigations, it is unknown whether the suspect is innocent or guilty. In laboratory studies, by contrast, it is known whether the suspect is innocent or guilty. Thus, in a laboratory study, the correct ID rate is the proportion of guilty suspects identified from the target-present lineups. The false ID rate is the proportion of innocent suspects identified from the target-absent lineups. Correct ID rates and false ID rates are then used to assess the discriminability of a lineup procedure.

<Table 1 near here>

There are different approaches to obtaining the false ID rate. First, an individual could be designated as the innocent suspect, and thus computing the false ID rate is straightforward (i.e., the number of times the innocent individual was identified divided by the number of target-absent lineups). Often, there is no actual innocent suspect in a lab study, and there are several different approaches to computing the false ID rate. One
approach is to randomly designate one filler in a target-absent lineup to serve as the innocent suspect after data collection (such that any identification of that individual would count as an innocent suspect ID). A second approach is to designate the filler who is most often misidentified as the innocent suspect (and any identification of that individual would count as an innocent suspect ID). A third approach is to estimate the number of innocent suspect identifications from the number of filler IDs from target-absent lineups. This estimate is obtained by dividing the number of filler IDs by the number of lineup members (e.g., six for the lineup in Figure 2). That estimated value is then divided by the number of target-absent lineups to estimate the false ID rate. Depending on the fairness of the lineup, those approaches should yield similar conclusions.

2.2 Discriminability

2.2.1 Past measure of discriminability: The diagnosticity ratio

For many years, claims about discriminability were almost always made on the basis of the diagnosticity ratio. The diagnosticity ratio is a likelihood ratio in which the correct ID rate is divided by the false ID rate. For example, based on the higher diagnosticity ratio that was frequently observed for the sequential lineup procedure, that procedure was declared diagnostically superior to the simultaneous procedure (Steblay, Dysart, Fulero, & Lindsay, 2001; Steblay, Dysart, & Wells, 2011). The same conclusion was reached when the simultaneous procedure was compared to the showup procedure, also using the diagnosticity ratio (Steblay, Dysart, Fulero, & Lindsay, 2003). However, the diagnosticity ratio is not the best way to measure discriminability because it conflates response bias, which is the likelihood of picking someone from a lineup, with discriminability (Wixted & Mickes, 2012; Gronlund, Wixted, & Mickes, 2014). Instead, Receiver Operating
Characteristic (ROC) analysis is a more suitable measure of discriminability (National Research Council, 2014).

2.2.2 Non-parametric measure of discriminability: Receiver operating characteristic (ROC) analysis

ROC analysis is widely used in other fields, particularly in applied fields, such as diagnostic medicine (Lusted, 1971a, 1971b; Swets, 1988; Swets, Dawes, & Monahan, 2000). In medicine, ROC analysis has been the preferred methodology since the 1970s for identifying the most accurate procedure for identifying a disease in a patient. The goal in diagnostic medicine is always to identify the procedure that yields the best discriminability (in this case, the ability of a medical procedure to distinguish between those who have the disease from those who do not). To maximize discriminability, the best procedure is the one that reduces both errors that can occur: 1) incorrectly identifying a disease in a patient without the disease, and 2) failing to identify the disease in a patient who has the disease. The former is analogous to incorrectly identifying an innocent suspect in a lineup or showup (a false ID), and the latter is analogous to not identifying the guilty suspect in a lineup or showup (a miss). Because identifying a disease in a patient is conceptually the same as identifying the guilty suspect in a lineup, ROC analysis was adapted for use in lineup data (Wixted & Mickes, 2012; Mickes, Flowe & Wixted, 2012).

An ROC plot is a plot of the correct ID rate and false ID rate for every level of response bias (confidence) and is shown using hypothetical data in Figure 3. The corresponding data used to plot the ROC curves are shown in the table below in red font. The bolded values represent the overall correct ID rates and false ID rates for both procedures (and the values used to compute diagnosticity ratios). In Figure 3, the leftmost
point is the correct ID rate and false ID rate reflecting those identifications made with the highest level of confidence (100% confident, the most conservative response bias). The next point to the right is the correct and false ID rate pair for those identifications made with medium confidence and above. The rightmost point is the correct rate and false ID rate for those identifications made with low, medium or high levels of confidence. In other words, the rightmost point is the overall correct ID rate and false ID rate. The dashed line is the line of chance performance, and if the points fell on that line it would mean that the eyewitness has no ability to distinguish innocent from guilty suspects. The further the curve falls above the line of chance performance, the better the discriminability. That is, the better eyewitnesses can distinguish innocent from guilty suspects. In Figure 3, it is clear that Procedure A gives rise to greater discriminability than Procedure B, because the ROC is further from the line of chance performance. To measure the statistical differences between the curves, the partial area under the curve (pAUC) is computed and compared (as described in detail in Gronlund, Wixted, & Mickes, 2014).

<Figure 3 near here>

To conduct ROC analysis, one must collect data from hundreds to thousands of participants per condition because there is only one trial per participant (just like the experience of a real eyewitness). Despite lineup data consisting of one decision per eyewitness, the combined dataset, which contains individuals with different levels of discriminability and different criterion settings, can nevertheless be used to assess discriminability. No additional assumptions need to be made about the form of the underlying memory distributions or characteristics of the decision process.
2.2.2.1 Arguments regarding ROC Analysis

The US National Academy of Sciences committee commissioned to assess the current state of eyewitness identification research made the following statement about ROC analysis:

“ROC analysis is a positive and promising step, with numerous advantages. For example, the area under the ROC curve is a single-number index of discriminability. Moreover, this index reflects a parameter-free approach to binary classification performance; the outcome is entirely data-dependent and thus identical across all users drawing from the same data set (Green & Swets, 2009). Most importantly for its application to the problem of evaluating eyewitness performance, the ROC approach possesses a distinct advantage because the dimensions of analysis – discriminability and response bias – map directly onto the mechanistic parameters of causal models of human recognition memory (see Chapter 4).” National Research Council, 2014; p. 59)

Despite the committee’s support of ROC analysis for lineup data, its use is not without its critics (Lampinen, 2016; Wells, Smalarz, & Smith, 2015). Wells et al. claimed that: “Yes, the National Research Council (NRC) report got it wrong by interpreting ROC analyses on lineups as measures of underlying discriminability,” (p. 325, Wells, Smith, & Smalarz, 2015). Each of their arguments is described below (and discussed in detail in Wixted & Mickes, 2015a; Wixted & Mickes, 2015b).

2.2.2.1.1 Anti-ROC argument #1: Eyewitness identification and diagnostic medicine ROCs are different

The first argument is that eyewitness identification ROCs and diagnostic medicine ROCs are different. The fact that lineups have fillers, and medical tests do not, may be considered as problematic. As mentioned previously, ROC analysis is the preferred method of analysis in medicine when the question is which diagnostic test is better at discriminating disease from non-disease states (e.g., Zweig & Campbell, 1993). Lineups are
conceptually analogous when the question is which lineup procedure, for example, is better at discriminating guilty from innocent suspects (Mickes et al. 2012). Although arguments against this claim have been advanced (Wells, Smalarz, & Smith, 2015), it is clear the case in favor of the similarity between ROCs from diagnostic medicine and ROCs from eyewitness identification is compelling. Consider, for example, detection-plus-quadrant-localization tasks in which radiologists are presented with an x-ray sectioned in four quadrants and their task is to identify the location of a tumor, if present, from one of four quadrants. This is isomorphic to an eyewitness being asked to identify the location of a perpetrator, if present, in one of six positions of a lineup. In both the x-ray and the lineup, the participant is presented with a target-present (tumor or guilty suspect present) or target-absent (no tumor or no guilty suspect present).

The top panel of Figure 4 (from Wixted & Mickes, 2015a, Figure 2) shows ROC curves constructed using two approaches with the data reported by Starr, Metz, Lusted, and Goodenough (1975). One approach considers all identifications made from a target-present x-ray (this includes correct IDs of the tumor and “filler” IDs, which are identifications made to any quadrant), with the corresponding rates plotted on the y-axis. All identifications (to any quadrant) made in target-absent x-rays, and the corresponding rates, are plotted on the x-axis. This approach yields the ROC curve labeled “ROC”. The second approach gives rise to location ROC (LROC), and includes only tumor IDs from the quadrant in which the tumor is located as correct IDs. This approach is analogous to lineup ROCs (which only count guilty suspect IDs from target-present lineups when computing the correct ID rate).

<Figure 4 near here>
As a comparison, the ROC and LROC plot from the Starr et al. (1975) study, and the ROC and LROC from an eyewitness study (Palmer, Brewer, Weber, & Nagesh, 2013), are placed side by side. In the Palmer et al. study, the ROC curve includes all identifications from target-present and target-absent lineups. The LROC curve only includes correct IDs made from target-present lineups. The false alarm rate is computed from all target-absent filler IDs (bottom horizontal axis) and the top horizontal axis shows the estimated false ID rate (by dividing the x-axis values by lineup size). Thus, the fact that the LROC curve and the ROC curve are the same whether filler IDs are included is evidence that including filler IDs does not change conclusions. The fact that the two plots are undeniably similar underscores the point that diagnostic medicine ROCs and eyewitness identification ROCs are inherently alike.

2.2.2.1.2 Anti-ROC argument #2: ROC analysis ignores filler identifications

The second argument is that ROC analysis needs to include filler identifications to meaningfully measure discriminability. In the past, identifications of fillers (also known as “foils”) were not considered of importance in the applied sense. Wells & Turtle (1986) made this point:

“Note, however, that the identification of a foil, although a true identification error, is a "known" error. That is, in an actual lineup or photo-spread situation, the identification of a foil will be detected as an error. The identification of a foil does not result in charges being brought against the identified person. In other words, the identification of a foil is not a false identification in the forensic sense. Throughout the remainder of this article, we use the term false identification only to refer to the identification of an innocent suspect; we call inaccurate identifications of foils foil identifications.” (emphasis added; p. 321)

Since the importation of ROC analysis for lineup data, however, there has been much discussion about filler identifications. Wells, Yang, and Smalarz (2015) claimed that ROC analysis is problematic because identifications made to fillers are not treated as errors. A
new term was then introduced, filler syphoning, that describes the idea that innocent suspects are less likely to be chosen (and therefore protected) because fillers are chosen instead (Wells, Smalarz, & Smith, 2015). They argue that some of the differences across ID procedures, for example, occur due to differential filler syphoning, and not to differential discriminability. They further argue that ROC analysis masks filler syphoning because it ignores filler identifications. There are three major challenges to that argument. First, filler identifications were not included in the diagnosticity ratio either, making ROC analysis no different in that regard. Second, as indicated above by Wells and Turtle (1986), only the identifications of guilty and innocent suspects matter, not filler identifications. Third, filler identifications can be included in ROC analysis, and the conclusions do not change.

There is an important difference between diagnostic x-ray ROCs and lineup ROCs, and it speaks to the irrelevance of filler IDs in lineup ROCs. In diagnostic medicine, “filler” IDs (i.e., deciding that a tumor is present in a quadrant that does not include a tumor) is an unknown error. That is, a doctor will need to perform an additional test to determine whether the quadrant he or she chose should be classified as a hit (tumor present), a false alarm (no tumor present in that quadrant), or even a false alarm and a miss (no tumor present in the chosen quadrant but a tumor was present in an adjacent quadrant). However, the police need not conduct any additional test when a filler is chosen; they know this individual to be innocent. A strict analog of the x-ray and lineup situations would occur only when prior knowledge exists of where a tumor is located (as would be the case when a senior doctor, who knows the answer, is training a junior doctor who does not). Only in this situation would the senior doctor know to ignore the ID from a “filler” quadrant, just like the police already know regarding the ID of a filler.
2.2.2.1.3 Anti-ROC argument #3: The diagnosticity ratio is what the legal system wants to know

The third argument is that the diagnosticity ratio is what the legal system really wants to know. However, results from both the diagnosticity ratio and ROC analysis are useful for the legal system, and which analysis should be considered depends on the question of interest because the two analyses address distinct questions. When trying to answer the question: what is the probability that a particular suspect who has been identified from a lineup is guilty, the diagnosticity ratio, not ROC analysis, is useful (Mickes, 2016; Mickes, Flowe, & Wixted, 2012). But when trying to answer the question: which procedure is diagnostically superior, ROC analysis, not the diagnosticity ratio, is required. This distinction is new to eyewitness ID research, but the distinction is well understood in medicine. Policymakers and leaders of health organizations (e.g., American Diabetic Society) endorse the test that yields the best discriminability based on results from ROC analysis. Medical doctors and their patients, on the other hand, need to know for a given diagnosis, how likely is it that they have the disease (e.g., diabetes). In this case, results from a likelihood ratio are relevant. Failing to understand this distinction, and consequently using the diagnosticity ratio to make conclusions about discriminability, has resulted in misleading conclusions.

2.2.2.1.4 Anti-ROC argument #4: ROC analysis measures “psychological” discriminability, not objective discriminability

The fourth argument is that ROC analysis measures “psychological” discriminability, not “objective discriminability”, but in fact ROC analysis can be used either for applied purposes or theoretical purposes. For applied purposes, ROC analysis assesses objective discriminability and merely graphically depicts the data. In this case, no
theoretical considerations are needed. Indeed, just like in diagnostic medicine, the typical use of ROC analysis in eyewitness identification is for applied purposes. Consider data from Wetmore et al. (2015) in which they compared performance on a showup (only a suspect, no fillers, are shown) to performance from a fair lineup.

To provide a concrete example, imagine 200 lineups and 200 showups. Figure 5 depicts identified guilty suspects (in black) and identified innocent suspects (in red) for both a showup (on the top left) and a lineup (on the top right). Considering only the fair lineups and showups (using the “innocent weak” suspect), Wetmore et al. found, for the showup, the overall correct ID rate was 61% and the overall false ID rate was 42%, and, for the lineup, the correct ID rate was 67% and the false ID rate was 10% (as shown in the top panels). No theoretical considerations are needed to appreciate that the lineup yields higher objective discriminability in that there are overall more guilty suspects and simultaneously fewer innocent suspects identified with the lineup compared to the showup. The figure below shows the Wetmore et al. (2015) data plotted in ROC space. The rightmost points outlined in red represent the overall correct and false ID rates for the showup and lineup. Whether considering only the most liberal point on the ROC (points outlined in red), or the entire ROC curves, the story is the same: the lineup is superior.

<Figure 5 near here>

ROC analysis can also be used for theoretical purposes to measure “psychological discriminability”, which is discriminability in the mind of a participant. This can be mimicked using a simple signal detection model (top panel of Figure 6) that maps onto the lineup situation (for a similar instantiation, see Clark, 2003). The model assumes that a
witness first determines the lineup member with the greatest memory strength, and then identifies that lineup member if his memory strength exceeds a response criterion, c (otherwise, the lineup is rejected). This model assumes Gaussian distributions of memory strength for fillers, innocent suspects, and guilty suspects with means of μFiller, μInnocent, and μGuilty, respectively, and standard deviations of 1 (assuming equal variance). A 6-member target-present lineup is conceptualized as 5 random draws from the Filler distribution and 1 random draw from the Guilty distribution. The conceptualization of a target-absent lineup depends on whether the lineup is fair (i.e., the suspect does not stand out amongst the fillers) or unfair. If the lineup is fair, the distribution for fillers and the innocent suspect are the same, and a 6-member target-absent lineup is conceptualized as 6 random draws from the Filler distribution. However, if the lineup is unfair, a 6-member target-absent lineup is conceptualized as 5 random draws from the Filler distribution and 1 random draw from the Innocent suspect distribution. In all cases, the model assumes that the eyewitness always selects the lineup member with the greatest associated memory strength provided that this strength exceeds a response criterion, c. The ability of eyewitnesses to discriminate innocent from guilty suspects is of interest, which is represented by the distance between the means of the μInnocent and μGuilty distributions (i.e., the greater the distance, the greater the discriminability).

Figure 6 shows two signal detection models. The model on the top left represents a fair lineup where the innocent suspect and the fillers resemble each other (i.e., μFiller = μInnocent) and the model on the bottom left represents an unfair lineup where the innocent suspect matches the guilty suspect much more than the fillers do (i.e., μFiller < μInnocent). The corresponding fair and unfair lineup ROC curves on the right show the
fact that the area under the curve is greater for the fair lineup than the unfair lineup.

Thus, ROC analyses can also be used to make sense of psychological discriminability (i.e., there are three distributions in an unfair lineup, in a fair lineup, the model can be reduced to a two distribution model). Although there are obvious parallels between the two types of discriminability, objective discriminability informs us about the accuracy of a procedure (and requires no theoretical assumptions) whereas psychological discriminability informs us about the capabilities of an individual, a conclusion that rests on several theoretical assumptions.

<Figure 6 near here>

2.2.3 Parametric measure of discriminability: $d'$

Because ROC analysis was only recently introduced to the eyewitness literature, many of the variables that have been investigated in the past, and which we review below, have yet to be examined using ROC analysis. That means that what has previously been declared a performance difference could in fact have affected response bias and not discriminability. However, $d'$ (a parametric measure of discriminability, i.e., the difference between $\mu_{\text{Innocent}}$ and $\mu_{\text{Guilty}}$ in the top panel of Figure 6) generally provides a closer approximation to the truth than the diagnosticity ratio in that $d'$ is little affected by response biases (Mickes, Moreland, Clark & Wixted, 2014). Figure 3 shows the suspect ID rates, filler ID rates, and no IDs for all levels of confidence for target-present and target-absent lineups for the two procedures in the ROC plot. All three summary statistics (the diagnosticity ratio, pAUC, and $d'$) are presented as well. In each case, Procedure A outperforms Procedure B.
The problem with using the diagnosticity ratio to measure discriminability can be easily illustrated. For each level of confidence, the diagnosticity ratio can be computed, so for Procedure A, the diagnosticity ratio for identifications made with low, medium, and high levels of confidence are 6.8 (.536/.079), 10.0 (.357/.036), and 22.0 (.157/.007), respectively. For Procedure B, the diagnosticity ratio for identifications made with low, medium, and high levels of confidence are 2.5 (.257/.103), 3.1 (.143/.046), and 4.0 (.057/.014), respectively. Thus, the diagnosticity ratio increases monotonically as responding becomes more conservative. It is an error to assume that the higher diagnosticity ratio is more accurate (i.e., results in better discriminability) because it is another way of saying that conservative responding is more accurate. In this example, regardless of response bias – liberal to conservative – discriminability is the same across the ROC curve. The $d'$ values, on the other hand, do not markedly change as responding becomes more conservative. For Procedure A, $d' = 1.4$ for identifications made with low confidence and $d' = 1.5$ for identifications made with medium and high levels of confidence. For Procedure B, $d' = 0.6$ for identifications made with low, medium, and high levels of confidence. The stability of $d'$ across changing response biases is why $d'$ is a more appropriate measure of discriminability than the diagnosticity ratio. Therefore, to assess the current state of knowledge (sections 3, 4 and 5), we compute $d'$ values to assess discriminability in the cases where ROC analysis has yet to be conducted.

2.3 Reliability

The reliability of an identification is assessed by considering positive predictive value. Positive predictive value is the probability that an identified individual is actually the perpetrator. The most effective way to measure positive predictive value is to take confidence at the time of identification into account. Highly confident witnesses testifying
in a court of law are persuasive to jurors (Penrod & Cutler, 1995), however, this confidence is not necessarily diagnostic of accuracy. There are many powerful forces that can adversely influence confidence, such as providing feedback immediately after an identification (e.g., Bradfield, Wells, & Olson, 2002; Wells & Bradfield, 1998), providing misinformation (e.g., Loftus, 2005), and making suggestive statements (e.g., Zaragoza & Lane, 1994; Lindsay & Johnson, 1989; Loftus & Palmer, 1974). Thus, confidence expressed at trial is analogous to considering evidence from a contaminated crime scene. However, confidence expressed at an initial identification attempt (assuming that the lineup and its administration are fair) is diagnostic of accuracy (Wixted, Mickes, Clark, Gronlund, & Roediger, 2015). However, social media platforms (e.g., Facebook, Instagram) have presented new ways for eyewitnesses to try to find the perpetrator, which may contaminate a subsequent identification procedure administered by the police (which is assumed to be the initial identification attempt). Consequently, the police should instruct witnesses not to search for faces on social media, just as they should instruct witnesses not to speak to one another (Technical Working Group for Eyewitness Evidence, 1999; see Paterson, Kemp, & Ng, 2011), and inquire if a witness did search for faces prior to administering the lineup test. To reiterate, it is initial confidence administered by the police that we refer to throughout this chapter. Moreover, we focus on the accuracy associated with high confidence IDs because the surprisingly high levels of accuracy associated with these high confidence IDs, and the capability of eyewitnesses to compensate for variables that adversely affect discriminability, is central to our claim that eyewitness evidence can be far more trustworthy than is typically thought.
2.3.1 *Past measure of the confidence-accuracy relationship: The point biserial correlation coefficient*

The confidence that an eyewitness expresses at initial identification, and the associated accuracy, were long thought to be weakly related. Early research yielded a point-biserial correlation coefficient of only .07, which prompted the following statement, “. . . the eyewitness confidence–accuracy relation is weak under good laboratory conditions and functionally useless in forensically representative settings” (p. 165, Wells & Murray, 1984). A decade later, that sentiment was echoed, “A major source of juror unreliability is their reliance on witness confidence, which 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, Penrod & Cutler, 1995). Sporer, Penrod, Read and Cutler (1995) conducted a meta-analysis that showed a relatively strong point bi-serial correlation coefficient of .41 when only choosers were included. Nonetheless, the overarching message was still to be cautious about confidence.

These misconceptions arose because a misleading statistic was used to assess the relationship between confidence and accuracy. Juslin et al. (1996) showed that the point-biserial correlation coefficient can mask the strength of the confidence-accuracy relationship, and argued that calibration analysis should be conducted instead. Since then, Brewer and colleagues have shown time and time again that the relationship is typically strong, when using the more appropriate calibration analysis (e.g., Brewer & Wells, 2006; Palmer, Brewer, Weber, & Nagesh, 2013; Sauer, Brewer, Zweck, & Weber, 2010), and when the focus is on only choosers (i.e., those who made an identification) at an initial identification attempt. Below, we discuss three ways to measure positive predictive value using confidence.
2.3.2 Calibration analysis

Calibration analysis measures the relationship between the subjective probability that an identification is correct and the objective probability that the identification is correct. There are different approaches to computing calibration, and one approach involves including the filler picks. Using that approach, to compute the dependent variable, the correct suspect IDs are divided by the sum of the correct IDs and incorrect IDs (including the filler IDs) for every level of confidence. That is then plotted as accuracy as a function of confidence. The top panel of Figure 7 shows the calibration plot of hypothetical data.

2.3.3 Confidence-accuracy calibration (CAC) analysis

A close relative of calibration analysis that only considers suspect identifications is confidence-accuracy calibration (CAC) analysis (Mickes, 2015). To compute the CAC dependent variable, the correct suspect IDs are divided by the sum of the correct suspect IDs and incorrect suspect IDs (fillers are excluded). One advantage CAC analysis has over calibration analysis is that the scale does not need to be from 0-100%, and instead confidence can be assessed using any type of ordinal scale. More importantly, CAC analysis specifically addresses the question that judges and jurors have (unlike calibration): What is the likelihood that this suspect (because only suspect IDs, not filler IDs, are advanced to trial) is guilty? The middle panel of Figure 7 shows the CAC plot of the same hypothetical data as in the top panel. Note that chance accuracy in a CAC plot is 50% if an equal number of target-present and target-absent lineups are used.

<Figure 7 near here>

Another way to measure positive predictive value is to compute the diagnosticity ratio for each level of confidence. As already stated, the diagnosticity ratio is not a
measure of discriminability, but it can be used to measure reliability. The bottom panel of Figure 7 shows the diagnosticity ratios by confidence for the hypothetical data used in the panels above. In all three panels of Figure 7, confidence is associated with higher levels of accuracy for Procedure A than Procedure B. Because judges and jurors are the stakeholders of this information, it might be easier to present CAC analysis to them because it is more understandable than a diagnosticity ratio. For example, saying that high confidence identifications are associated with 99% accuracy is more understandable than saying the diagnosticity ratio was 103.

We now have laid out how to measure discriminability (ROC analysis, or d’ if ROC analyses have not yet been conducted) and how to assess reliability by measuring the relationship between confidence and accuracy (CAC analyses). It is time to turn our attention to prior research involving the many variables that affect eyewitness memory. We organize this discussion by considering whether the primary impact of these variables takes place during the encoding, storage, or retrieval phases of memory.

3 Variables that Affect Encoding

3.1 Weapon Focus

Weapon focus refers to a reduction in discriminability for the perpetrator and the details surrounding a crime that involved a weapon (Loftus, 1979; Loftus, Loftus, & Messo, 1987). The weapon focus effect is a popular phenomenon to investigate. A search of the term “weapon focus” yields 1,530 results in Google Scholar (March 15, 2016). The effect likely occurs during encoding because attention is directed to the weapon rather than on the other aspects of the crime or the perpetrator. Two prominent theories, both proposed in the original empirical weapon focus studies, attempt to account for the effect and continue to dominate the literature. These accounts posit that weapon focus is due either
to high arousal or to the presence of an unusual object (Loftus et al.). Loftus et al.,
crediting Easterbrook (1959), proposed that the presence of a weapon causes high arousal
that might consequently narrow an eyewitness’ attentional focus. They also proposed that
lower accuracy might be due to a tendency to fixate on an unusual object. The two
possibilities are commonly referred to as the arousal/threat hypothesis and the unusual
item hypothesis, respectively. Since then, investigations have tested the predictions of
both hypotheses, with some support for each account.

Some of the studies test recall of events, and/or test recognition of the
perpetrator from a lineup. In many of the latter studies, memory was tested only on
target-present (e.g., Shaw & Slotnick, 2001; Pickel, 1998) or target-absent lineups (e.g.,
Hulse & Memon, 2006; Maass & Kohnken, 1989), which is problematic for measuring
discriminability (e.g., Mickes & Wixted, 2015; see Rotello, Heit, & Dube, 2015). The
conclusion of a meta-analysis of weapon focus effect studies was that the effect was
dependent on the level of threat (the larger the threat, the bigger the effect) and
retention intervals (the longer the retention interval, the smaller the effect).

The first weapon focus experiment in which ROC analysis was conducted was
recently published (Carlson & Carlson, 2014). In this experiment, participants watched a
video of a crime in which the perpetrator, who had a distinctive feature or not, had a
weapon or not, and then were tested on a sequential or simultaneous lineup.
Discriminability was lower in the weapon present condition than in the weapon absent
condition when the distinctive feature was not present. Thus, using ROC analysis, the
weapon focus effect was found.

The next question is: what is the effect of weapon focus on the reliability of an ID
(the relationship between confidence and accuracy as assessed by CAC analysis)?
Regardless of whether a weapon causes lower discriminability, what matters in the applied sense are, can eyewitnesses take a factor (like weapon presence) into account when reporting their level of confidence (i.e., the CAC results). These are the results that can inform judges and jurors when deciding about defendants’ culpability. If accuracy for identifications made with high confidence is the same regardless of the presence or absence of a weapon at encoding, then that is a more relevant consideration than if memory is overall worse if a weapon is present during the crime. We limited the CAC analysis to the weapon and no weapon conditions from Carlson and Carlson (2014) (and excluded the condition in which an artificial feature was added to the target). Our CAC analysis revealed that high-confidence accuracy was very high whether a weapon was present or not (e.g., average accuracy, collapsed across condition, was 97%).

In a more recent investigation of the reliability of the weapon focus effect, participants were assigned to one of three conditions: weapon present, weapon present but concealed, and weapon absent (Carlson, Dias, Weatherford, & Carlson, 2016). There was a large discriminability difference between the weapon present and weapon absent and concealed conditions, but the story was different for reliability (and replicated our CAC results of their earlier study). CAC analysis revealed no statistical differences among conditions for the identifications made with medium and high levels of confidence. If this result continues to replicate, these are the results that judges and jurors need to know: Not that the presence of a weapon may yield lower discriminability, because there is no control over whether a weapon was wielded or not during a crime, but the fact that at the high end of the confidence scale, the likelihood that the suspect is guilty (the reliability) is similar, and high, irrespective of the presence of a weapon.
3.2 Exposure Duration and Divided Attention

The same distinction between discriminability and reliability should be made when it comes to other variables that affect eyewitness identification, such as exposure duration, divided attention, etc. No one would dispute that longer exposure times to stimuli, including a perpetrator, should generally result in greater discriminability than shorter exposure times. Nor would anyone dispute that greater attention paid during encoding should generally lead to greater discriminability. Palmer et al. (2013) conducted experiments in which they manipulated two variables that affect encoding: exposure duration (5 seconds vs. 90 seconds; Experiment 1), attention (full vs. divided; Experiment 2), and one variable that affects retrieval: retention interval (immediately tested vs. tested after a delay; Experiment 1). In Experiment 1, one experimenter approached potential participants in public places and asked if they would take part in an experiment. If they agreed, a second experimenter would step into view for either 5 seconds or 90 seconds; the participants were tested immediately or 6-8 days later. Collapsing across the retention interval conditions, ROC analysis showed that the 90-second encoding condition yielded greater discriminability than the 5-second condition. What about the confidence-accuracy relationship? Palmer et al. provided the data to conduct CAC analysis so that we could answer the question: Did those in the 5-second condition appreciate the fact that they only saw the target briefly and adjust their confidence accordingly? Yes. Identifications made with high confidence had the same accuracy regardless of exposure duration (the CAC results of their experiment are reported in Mickes, 2015).

In Experiment 2, participants watched a video of two target individuals. Participants in the divided attention condition had to respond to low and high tones while watching the video. Those in the control condition were told to ignore the tones while
watching the video. As expected, discriminability was lower in the divided attention condition compared to the full attention condition. Importantly, similar to the exposure duration findings, participants were able to calibrate their confidence to reflect the fact that they were more likely to make an error in the divided attention condition. Identifications made with high confidence were highly accurate for both conditions (over 96%). Once again, participants can adjust their confidence to reflect the likelihood they are making an error with these variables that affect memory at encoding.

### 3.3 Cross-Race Bias

The cross-race bias is a phenomenon in which people are more accurate at recognizing faces of individuals of the same race than individuals of another race (Chance & Goldstein, 1981; Malpass & Kravitz, 1969; Meissner & Brigham, 2001). The lower discriminability (i.e., ability to discriminate old from new faces) for other race faces is generally indicated by an increased false alarm rate and occasionally by a reduced hit rate (Meissner et al., 2001). Participants also generally exhibit a more liberal response bias for cross-race faces, which means that participants are more likely to endorse cross-race faces as previously seen, irrespective of whether they actually were. The cross-race bias has been found across a variety of conditions and across a variety of races (e.g., Ng & Lindsay, 1994; Platz & Hosch, 1988; Wright, Boyd, & Tredoux, 2001).

The cross-race bias is thought to arise during encoding. Bornstein, Laub, Meissner, and Susa (2013) were unable to reduce the deficit by delivering cautionary instructions at retrieval. Golby, Gabrieli, Chiao, and Eberhardt (2001) used functional MRI and found greater fusiform face area activation, the first stage of face-specific processing, for same race faces. However, Johnson and Frederickson (2005) found that the induction of positive
emotions, either prior to or after encoding the same race and cross race faces, eliminated the cross-race deficit (by raising $d'$ for the cross-race faces).

Attempts at explaining the cross-race recognition deficit have taken two general approaches. One approach attributes the deficit to differential perceptual experience (e.g., differences in the amount or quality of contact, Brigham & Malpass, 1985; Malpass & Kravitz, 1969), which has resulted in a wide-range of suggestions for how differential perceptual experience could be manifested (e.g., Freeman, Pauker, & Sanchez, 2016). For example, DeGutis, Mercado, Wilmer, and Rosenblatt (2013) found that own-race faces were processed more holistically than cross-race faces, and linked this increased degree of holistic processing to the own race recognition advantage. Face space explanations (e.g., Valentine, 1991) documented a denser (i.e., more confusable) psychological representation of cross race faces, which would arise if participants focus on the wrong features for discriminating cross race faces (Papesh & Goldinger, 2010). Goldinger, He, and Papesh (2009) documented that participants made fewer and longer eye fixations to cross-race faces, and also reported that some participants exerted less cognitive effort (as indexed by pupil diameter) to cross race faces.

The second approach attributes the cross-race deficit to differential social categorization. Bernstein, Young, and Hugenberg (2007) demonstrated the power of social categorization to differentially influence memory for in-group (same-race) versus out-group (cross-race) faces. White participants viewed a series of photos of White faces on red and green backgrounds. Those participants were told that the photos on a red background went to their university (in-group) and those on a green background went to their rival university (out-group). Participants showed greater $d'$ for the in-group photos
A control group exhibited no memory difference for the photos ($d' = 1.08$ and $1.14$, for red and green backgrounds). Consequently, it appears that the cross-race effect can be considered a cross-category or an out-group effect.

The most comprehensive current theory, the Categorization-Individuation Model (Hugenberg, Young, Bernstein, & Sacco, 2010), ties these two approaches together by proposing that the ability to remember faces that belong to a different category/race is a function of three factors: social categorization, motivated individuation, and perceptual experience. The core of the model involves opposing processes that operate at encoding: Individuation acts to distinguish category members from one another (something that people do more naturally for own-group members), whereas categorization clusters category members along shared dimensions (e.g., race, Levin, 1996). According to this theory, the out-group bias arises from the tendency to attend to identity-diagnostic (individualistic) features of in-group individuals but category-diagnostic features (e.g., skin tone) of out-group individuals. This theory also provides explanations regarding factors that create, diminish, or eliminate, an out-group memory deficit.

One factor that results in out-group individuals becoming more salient/memorable is when those individuals are threatening. Ackerman et al. (2006) found that the cross-race bias was eliminated if the other race faces displayed angry emotions because threat induces individuation, rather than categorization. Likewise, Shriver and Hugenberg (2010) diminished the magnitude of the cross-race deficit if other race faces were attributed greater power or status (e.g., a prestigious occupational title like doctor) because threat and power enhance the likelihood of individuation, reduce reliance on categorization, and thereby enhanced recognition memory.
Hourihan, Benjamin, and Lui (2012) found that participants also are worse at making metamemory judgments to cross-race faces regarding whether they will subsequently remember the face. This has potential implications for the reliance on eyewitness confidence when a cross-race ID is involved. If eyewitnesses are not aware of this difficulty, they may not be able to adjust their confidence judgment accordingly. But CAC analyses conducted on Dodson and Dobolyi (2015) showed that confidence was very informative of accuracy but race was minimally important (the accuracy for high confidence same race and cross race IDs both exceed 95%). Again, these participants were able to adjust the likelihood that they made a correct ID despite the fact that they were more likely to make an error when making a cross race ID. A similar finding was reported by Nguyen, Pezdek, and Wixted (in press) (for experiments in which performance was above chance levels), who re-analyzed data from four cross-race face recognition experiments.

3.4 Stress and Arousal

Arousal signifies general physiological and psychological activation; stress arises from an imbalance between the physical and psychological demands and the ability to respond (Hoscheidt, LaBar, Ryan, Jacobs, & Nadel, 2014). Stress and arousal have complex effects on memory. For example, stress appears to impair memory for neutral information (Payne, Nadel, Allen, Thomas, & Jacobs, 2002), but enhance memory for emotionally arousing materials (Buchanan & Lovallo 2001). The research conducted on lineups focuses on discriminability; more research needs to be completed assessing positive predictive value. We can never makes crime less stressful, but we need to determine if eyewitnesses are able to compensate for that stress when they assess their confidence.
A meta-analysis by Deffenbacher, Bornstein, Penrod, and McGorty (2004) concluded that high levels of stress adversely impacts eyewitness IDs of a target. Across 27 tests, correct ID accuracy was .42 versus .54 for high stress and low stress, respectively. Interestingly, they found that increased stress adversely affected target-present accuracy, not target-absent accuracy. Valentine and Mesout (2008) tested individuals visiting the London Dungeon who were exposed to an individual during the tour that they later were asked to ID from a target-present lineup. Valentine and Mesout did not manipulate stress but rather classified participants based on a median split based on their anxiety score in reaction to the Dungeon performance. Among the low anxious individuals, 75% chose the target, but only 18% of the high anxious individuals chose the target. Unfortunately, without a target-absent lineup for comparison, and given large differences in response biases across the two groups (high anxious more conservative), conclusions are tentative.

Morgan et al. (2004) conducted the most compelling experiment examining memory for a perpetrator. It is the most compelling because it involves far greater levels of stress than can ethically be used in the laboratory. Participants were soldiers undergoing prisoner-of-war training, which involved sleep and food deprivation. After a week of classroom training, participants were confined in a mock prisoner of war camp, and while in isolation were subjected to low-stress and high-stress interrogations from different interrogators. Participants were threatened with physical violence in the high-stress interrogation. ID attempts were made approximately 24 hours after completion of the training. Despite interacting with their interrogators for more than 30 minutes, participants had great difficulty subsequently identifying an interrogator (some viewed a live lineup, others a photo lineup). We took the simple average of the live and photo lineup data, and found that $d'$ was greater (and the criterion more liberal) when
identifying the low-stress interrogator ($d' = 2.33, c = 0.66$) than the high-stress interrogator ($d' = 1.36, c = 1.20$). Morgan et al. (2013) used a similar sample of participants undergoing prisoner-of-war training. All participants experienced high stress interrogations but varied in whether, and how, they were exposed to misinformation. Some groups subsequently viewed a 9-person target-absent lineup, from which 53% of the control participants made a false positive selection. Those individuals in the photographic misinformation condition (who were exposed to a filler photo and asked a series of questions about the interrogation while viewing this photo) had a false ID rate of 91%. Not surprisingly, the majority of these selections involved selection of the filler.

In a recent investigation of the effect of stress on lineup performance, participants watched a video of a mock crime while either being exposed to a stressor (cold pressor) or not, and later memory for the target in the video was tested from a lineup (Sauerland et al. 2016). Cortisol levels validated the efficacy of the stress manipulation, yet neither ROC analysis nor CAC analysis yielded a difference between the groups. But interestingly, identifications made with high confidence were high in accuracy (i.e., over 95% proportion correct) regardless of level of stress. However, the authors acknowledged that the sample size was small, so more research still needs to be conducted in this area.

In the Houston Police Department field study, over 300 real eyewitnesses to robberies who were faced with a lineup decision made confidence judgments in their identifications (Wells, 2014). Many of these witnesses were victims and thus likely had experienced some degree of stress during encoding of the crime. Despite that, the identifications made with high confidence were associated with high accuracy and identifications made with low accuracy were associated with low accuracy (Wixted,
Mickes, Dunn, Clark & Wells, 2016). These results suggest that these eyewitnesses were highly reliable.

In sum, stress appears to adversely impact discriminability for the target of an event, be that a mock perpetrator, an actor in the London Dungeon, or an interrogator. If the police could control the level of stress experienced by eyewitnesses at the time of a crime, then they should ensure low stress to ensure higher discriminability during a later identification. But of course they cannot, so the key question for the legal system concerns the effect of stress on reliability. Initial CAC results and implications of the Houston Police Department field study suggest that stress does not impair reliability (i.e., stress does not cause eyewitnesses to mistakenly identify innocent suspects with high confidence). If confirmed by additional research, this would be a fact that judges and jurors should be made aware of.

4 Variables that Affect Storage

4.1 Verbal Overshadowing

The verbal overshadowing effect posits that memory is adversely affected after providing a verbal description of a previously presented stimulus (e.g., a face). Schooler and Engstler-Schooler (1990) coined the term based on results from several experiments in which participants viewed a video of a mock crime. They found that those who verbally described the perpetrator were less able to correctly identify the perpetrator from a later lineup test than those in the control condition. In Experiment 1 of the original paper, participants were tested on target-present lineup procedures, and the correct ID rate was .64 in the control condition compared to only .37 in the verbal description condition. This counterintuitive finding sparked much interest and follow-up research: the original paper
has been cited 763 times (Google Scholar search retrieved April 8, 2016). Note that no conclusion can be reached regarding discriminability without a target-absent condition.

The follow-up research does not tell a consistent story. The lack of consistency was the impetus for a meta-analysis conducted by Meissner and Brigham (2001). Based on results from 29 investigations of the verbal overshadowing effect the authors concluded that the effect was real, but small. Other investigations were conducted on list learning experiments in which the verbalization conditions fared better than the control conditions. The change in paradigms (i.e., list learning vs. forensically relevant experiments) might account for the inconsistencies reported since the original verbal overshadowing effect paper (which used a forensically relevant one-trial paradigm).

There are three main hypotheses to explain the verbal overshadowing effect. The content account (e.g., Meissner et al., 2001) holds that the verbal description interferes with the memory of the target, causing a reduction in discriminability (i.e., the ability to distinguish between the innocent and guilty suspect). The criterion-shift account (Clare & Lewandowsky, 2004) holds that verbal overshadowing reflects a change in response bias (i.e., the likelihood to choose someone from a lineup) rather than a change in discriminability. The processing account holds that the switch from visual to verbal processing (Schooler, 2002) affects both discriminability and response bias (Chin & Schooler, 2008).

Because of the lack of consistent findings, and because the importance of conducting direct replications has recently been highlighted in the field of psychology (Pashler & Wagenmakers, 2012), Experiment 1 and Experiment 4 in the original Schooler and Engstler-Schooler (1990) paper were chosen as one of the first pre-registered replication report projects (Open Science Collaboration, 2015). Independent laboratories
responded to the call to attempt to replicate the finding, and the meta-analysis from that work can inform us about the true size of the verbal overshadowing effect (Alogna et al., 2014). The results revealed a small, but significant, verbal overshadowing deficit, which was larger if the description task was separated in time from the original event.

In the original paper, participants were only tested on target-present lineups. But to get a complete picture of how verbalization affects participants’ ability to discriminate innocent from guilty suspects, one needs to also measure false ID rates (which are collected from target-absent lineups). It is well established that while the correct ID rate could be lower in one condition this does not mean that that condition is worse, per se, because the false ID rate could be lower in that condition as well. Despite replicating the original result (Alogna et al., 2014), because there was no way to measure false IDs, the effect verbalization has on identification performance (discriminability) remains unclear (Mickes, 2016; Mickes & Wixted, 2015; Rotello et al., 2015).

Recently, two experiments investigating the effects of verbal overshadowing were completed that allowed ROC and CAC analyses (Seale-Carlisle & Mickes, 2015). Although the original and replication studies used 8-person lineups, we reduced it to 6-person lineups. In both experiments, the reduced correct ID rates replicated. In the experiment in which the description occurred immediately after the study phase, the ROC curves were not significantly different (i.e., discriminability was not different). CAC analysis revealed that identifications made with high confidence were equally reliable as well. In the experiment in which the description occurred 20 minutes after the study phase, and immediately before the test phase, the ROC curves revealed a difference. That is, discriminability was lower in the verbalization condition. However, CAC analysis revealed that high confidence identifications still were equally reliable, and both highly accurate.
The CAC results are reassuring given that the criminal justice system relies on verbal reports of crimes. Why these differences exist due to the timing of the descriptions should be the subject of future investigations (Mickes, 2016).

### 4.2 Retention Interval

What do judges and jurors need to know about increasing retention interval that they do not already know? No surprise that memory gets worse with time. But what about reliability? We suspect that judges and jurors have the wrong idea here. To make the point that judges and jurors need to know about only suspect identification accuracy, Wixted, Read, and Lindsay (2016) reanalyzed data from four retention interval studies. In each of the studies either calibration analysis or correlation coefficients were originally reported. Retention intervals varied and participants were tested either immediately or one week after the encoding event (Juslin et al., 1996; Palmer et al. 2013), immediately or three weeks after (Sauer et al. 2010), and 3 months or 6-to-9 months later (Read et al. 1998). Remarkably, the CAC results showed that regardless of retention interval, identifications made with high confidence were highly accurate, even after six months. In each case, discriminability (as measured by $d'$) decreased as retention interval increased. Thus, despite ever-decreasing discriminability with time, high confidence identifications remain highly accurate. This is a consideration judges and jurors should take into account.

### 4.3 Suggestibility Effects

It is well-known that memory is reconstructive (e.g., Bartlett, 1932; Roediger & McDermott, 1995), and that post-event suggestions are often incorporated and reported as part of an original event (e.g., Loftus & Palmer, 1974). This occurs, in part, due to poor source monitoring (Johnson, Hashtroudi, & Lindsay, 1993). Eyewitnesses become confused between their own experiences and inferences they draw (Do I remember the
tattoo or did I infer that the gang member who robbed me had a tattoo?), between what they experience and what a co-witness might report (Did you see the scar on the robber’s face?), or what they might see or hear in the media.

Bonham and González-Vallejo (2009) examined the effect of misinformation on eyewitness discriminability and calibration. Participants watched a mock crime video and then responded to questionnaires or read narratives that contained correct information and misinformation about what transpired in the video. Discriminability was poorer for details regarding the crime when misinformation was introduced. Likewise, the confidence-accuracy relationship suffered as a function of misinformation. In other words, individuals incorrectly maintained high confidence in the accuracy of memories about which they had been misinformed. Unlike factors like cross-race and retention interval, eyewitnesses appear to be unable to adjust their confidence to reflect the reduced accuracy that arises from misinformation.

5 Variables that Affect Retrieval

5.1 Creating Fair Lineups

What constitutes a fair lineup and how a fair lineup is created has much to do with the fillers selected to be in that lineup. In fact, Woltager, Malpass, and McQuiston (2004) argued that part of the reason why there are fillers is to protect the innocent suspect from being wrongfully identified.

5.1.1 Filler Selection

What should the characteristics of the fillers be? Should the fillers in the lineup match the verbal description of the perpetrator or should they match how the suspect (who may be innocent or guilty) looks? And how closely should the fillers match? How
should perpetrators with distinctive characteristics be matched? These are questions that have been the focus of many investigations, because knowing the answers to these questions can result in the preparation of the fairest lineup and the resulting best quality eyewitness evidence.

5.1.1.1 Description-matched or Suspect-matched?

Wells, Rydell and Seelau (1993) tested the hypothesis that the description-matched method of filler selection would result in greater discriminability than the suspect-matched method of filler selection. Indeed, the description-matched condition yielded a higher $d'$ than the suspect-matched condition (1.61 vs. 0.37, respectively). They concluded, “A good lineup appears to be one in which all lineup members match the eyewitness’s pre-lineup description of the culprit but otherwise do not resemble each other.” (p. 844). Although follow-up research has replicated that finding (e.g., Juslin et al. 1996, Tunnicliff & Clark, 2000), other findings support greater discriminability for the suspect matched conditions (e.g., Lindsay, Martin & Webber, 1994; Tunnicliff & Clark, 2000; Darling, Valentine, & Memon, 2007). Clark, Rush, and Moreland (2013) found no evidence that description-matched filler selection is better, and recommended a combination of description- and suspect-matched selection, which is what many police departments appear to do (Wise, Safer, & Maro, 2011). More research is needed regarding how best to select fillers for a lineup, and how similar those fillers should be to a suspect.

5.1.1.2 Distinctive Features

Distinctive features, such as a face tattoo or scars, may be encoded during the crime, but how those features are subsequently handled is an issue for law enforcement.
There are several propositions regarding distinctive features when assembling a lineup: to conceal, duplicate across members, block over, or to leave alone. Currently, guidelines leave it to the discretion of the identification officer whether they replicate or conceal the feature (PACE Code D; Technical Working Group for Eyewitness Evidence, 1999).

In a recent comparison of replication, concealment, blocking or leave alone, data were collected from 9841 participants (Colloff, Wade, & Strange, in press). Participants viewed video of a crime (one of four videos with four different targets) in which there was a prominent distinctive feature (each target had a different distinctive feature). ROC analysis revealed that discriminability was similar for the replication, concealment and blocking conditions, but significantly lower for the leave alone condition. This finding implies that when constructing the lineup, ignoring the suspect’s distinctive feature is bad practice and any other alternative would be preferred.

The other alternatives also lead to better reliability. At all levels of confidence, accuracy was significantly lower in the leave alone condition. Moreover, at high levels of confidence (i.e., identifications made with 90-100% confidence), accuracy in the leave alone condition was only approximately 60%, whereas in the other conditions, accuracy was approximately 85%. This finding implies that judges and jurors need to know how the lineup was constructed. For example, if the defendant was identified with high confidence from a lineup in which there was a suspect with a distinctive feature that was not accounted for by concealing, replicating or blocking, then even an identification made with high confidence is less trustworthy.
5.2 Identification Procedure/Presentation

5.2.1 Simultaneous vs. sequential lineups

How should lineup members be presented to the eyewitness? In a simultaneous lineup, all lineup members are presented at once and only one decision (Which person, if anyone, is the perpetrator?). Lineup members are viewed one at a time in a sequential lineup, and a decision may be required regarding lineup member 1 (Is this the perpetrator?) before lineup member 2 is presented (although there are several variations on how the sequential procedure is conducted). In the first investigation comparing simultaneous versus sequential photo lineups (Lindsay & Wells, 1985), the false ID rate was much higher in simultaneous lineups than sequential lineups (.43 vs. .17, respectively). The correct ID rate was not much higher in simultaneous lineups than sequential lineups (.58 vs. .50, respectively). Over time, with the exception of two experiments (Lindsay, Lea, & Fulford, 1991a; Lindsay et al., 1991b), the differences in false ID rates were not as high as in the original study, and the sequential advantage sometimes failed to replicate (e.g., Carlson, Gronlund, & Clark, 2008; Gronlund, Carlson, Dailey, & Goodsell, 2009). Others concluded that the pattern of results was more consistent with the sequential lineup inducing a conservative criterion shift (e.g., Ebbeson & Flowe, 2002; Meissner, Tredoux, Parker, & MacLin, 2005; Palmer & Brewer, 2012). Despite this contrary evidence, the claim made was that the sequential lineup was superior (Steblay, Dysart, Fulero, & Lindsay, 2001; Steblay, Dysart, & Wells, 2011). The claims were based on higher diagnosticity ratios for the sequential lineups, and because of these claims, sequential lineups were recommended for use over simultaneous lineups (e.g., Lindsay, 1999; Innocence Project, 2009; Wells et al., 2000). As a result, 30% of jurisdictions across the US
switched from using the simultaneous lineups to sequential lineups (Police Executive Research Forum, 2013).

After the importation of ROC analysis to use for lineup data in 2012, the simultaneous lineup has been found to consistently outperform sequential lineups (e.g., Carlson & Carlson, 2014; Dobolyi & Dodson, 2013; Gronlund et al. 2012; Mickes, Flowe & Wixted, 2012). The US National Academy of Sciences committee decided not to endorse one procedure over the other because they deemed the matter still unresolved (National Research Council, 2014). Nonetheless, the Innocence Project ceased to support sequential lineups as a reform (http://www.innocenceproject.org/following-the-science/; retrieved May 15, 2016), although some US state Innocence Projects continue to support sequential lineups (e.g., Minnesota: http://ipmn.org/wp-content/uploads/2016/09/Minnesota-Eyewitness-Identification-Standard-Protocols.pdf, retrieved October 18, 2016). Two recent field studies have found that simultaneous lineups yield better discriminability (Amendola & Wixted, 2015; Wixted, Mickes, Dunn, Clark, & W. Wells, 2015), which further bolsters support for simultaneous lineups. In fact, some recent guidelines have stated that simultaneous lineups should be used. For example, the Pennsylvania police may follow guidelines that, “… recommend that officers present witnesses with photo arrays, showing multiple photographs of potential suspects, instead of one by one.” (http://www.mcall.com/news/nationworld/pennsylvania/mc-pa-new-guidelines-false-ids-20160412-story.html; retrieved April 12, 2016). Thus, it appears as though the tide is turning in favor of the simultaneous lineup, and if the results continue to replicate, then more policymakers should begin recommending the use of simultaneous lineups.

Regardless of the lineup procedure that is used, judges and jurors need to know about the reliability of an identification. This issue has been addressed by conducting CAC
analyses on experiments that reported ROC analyses. It has been shown that simultaneous and sequential lineups yield equal reliability for identifications made with high confidence (Dobolyi & Dodson, 2013; Gronlund et al. 2012; Experiments 1A and 1B of Mickes, Flowe & Wixted, 2012; Weber & Brewer, 2004). Thus, while discriminability suffers due to sequential lineup presentation, reliability (at the highest confidence level) is not different.

5.2.2 Showup Procedures

Unlike a lineup, a showup involves the presentation of a single suspect (not accompanied by fillers) to the eyewitness. Showups have been criticized for their inherently suggestive nature because the eyewitnesses obviously know that the person presented is the police suspect (Goodsell, Wetmore, Neuschatz, & Gronlund, 2013; Steblay, Dysart, Fulero, & Lindsay, 2003). Despite this, and other potential downsides, the showup will likely remain a common way of testing eyewitness memory given that it is an easy way to test memory soon after a crime has occurred.

The simultaneous procedure yields greater discriminability than the showup procedure when measured with ROC analysis (Wetmore, et al. 2015; Mickes, 2015; Gronlund et al., 2012). The CAC results differ as well. Showup procedures yield lower reliability even for the identifications made with high confidence compared to simultaneous lineups (Wetmore, et al. 2015; Mickes, 2015). Thus, there is a growing body of empirical evidence suggesting that simultaneous lineups are superior to showups in terms of both discriminability and reliability. This suggests that lineups should be used when possible if the results continue to replicate.
5.3 Blind Administration

The lineup administrator may influence an eyewitness in two ways. 1) The administrator can influence who an eyewitness chooses, or if an eyewitness chooses. 2) The administrator can provide feedback about a choice that an eyewitness makes. The first source of influence maps onto discriminability (who) and bias (if), respectively. A discriminability influence results from what Greathouse and Kovera (2009) referred to as steering, whereby the administrator directs a witness toward a particular suspect. Alternatively, bias is affected if an administrator exerts a non-specific influence that affects the rate at which an eyewitness chooses. There is evidence in the literature for both these patterns (Clark, Marshall, & Rosenthal, 2009; Greathouse & Kovera, 2009; Haw & Fisher, 2004). The second source of influence arises once an eyewitness has made a choice, and the administrator provides post-ID feedback regarding the correctness of that choice (e.g., Wells & Bradfield, 1998; Wright & Skagerberg, 2007; see meta-analysis by Bradfield Douglass & Steblay, 2006). Post-ID feedback tends to inflate eyewitness confidence, as well as inflate several other indices reflecting an eyewitness’ memory for the perpetrator (e.g., how long a look they got, how close they were). But if we follow Wixted et al.’s (2015) prescription to focus on the reports from the initial identification, post-ID feedback need not be of great concern. But double-blind administration remains important so that the reported confidence associated with the first, fair test of an eyewitness’ memory, remains unspoiled.

What do we know about the effects of blind administration on discriminability and response bias? Greathouse and Kovera (2009) varied lineup presentation, target presence, single- or double-blind, and biased or unbiased lineup instructions. We focus on the simultaneous lineup data, and computed $d'$ and $c$ from the reported correct and false IDs
rates. If biased instructions (which implied the suspect was in the lineup and the witness should identify him) were employed, discriminability was greater with double-blind than single-blind administration ($d' = 2.19$ vs. $d' = 1.52$, respectively), and double-blind administration induced greater response conservatism ($c = 0.74$ vs. $c = -0.32$, respectively). However, for unbiased instructions (the suspect may or may not be in the lineup), discriminability was slightly greater for single-blind than double-blind administration ($d' = 1.01$ vs. $d' = 0.70$, respectively), and no difference arose in response bias ($c = 0.58$ vs. $c = 0.53$, respectively). These data suggest that combining double-blind control with unbiased instructions might not be beneficial.

The National Academy of Sciences report (National Research Council, 2014) recommends double-blind lineup administration, and a long history of research in psychology and medicine support the merits of double-blind testing (Rosenthal, 2002). But there is limited research within the eyewitness domain that compares double- to single-blind lineup administration using target-present and target-absent lineups. Clark, Benjamin, Wixted, Mickes, and Gronlund (2015) provide a recent review of blind administration, and also mention the relative lack of empirical data that focuses on the crucial independent variables (blind vs. non-blind, both target-present and target-absent, without contamination from factors like different types of instructions, lineup presentation methods). We also note that more studies need to be conducted using skilled lineup administrators (e.g., Clark et al., 2013). Administrator effects are clearly going to be influenced by the skill (or lack thereof) of the administrators involved (Russano, Dickinson, Greathouse, & Kovera, 2006).
Although the adoption of double-blind lineups might seem like a no-brainer, the effects of double-blind lineup administration are complex. Clark et al. (2009) used lineup administrators who were blind to the presence or position of a suspect. The lineup administrator remained silent in the no-influence condition. In the subtle-influence condition, the administrator made innocuous statements like, “Take your time.” In the similarity-influence condition, the administrator asked if any lineup member resembled the perpetrator. For the data involving the designated innocent suspect, discriminability was better in the similarity influence ($d’ = 2.41$) than the no influence ($d’ = 1.67$) and subtle influence ($d’ = 0.68$) conditions. Clark et al. (2013) trained administrators to use techniques to influence witnesses. These administrators knew the position of the suspect, although the suspect was only guilty half the time, unbeknownst to the administrator. Surprisingly, Clark et al. found slightly higher $d’$ in the influence condition, likely because it is easier to steer a witness to a guilty suspect than to an innocent suspect.

In sum, there is evidence that blind administration can sometimes harm discriminability, and conversely, that non-blind administration can sometimes enhance discriminability. Of course, as Clark et al. (2015) argued, that does not mean that non-blind administration is to be recommended. Although a key goal of eyewitness reforms is to determine which procedures can enhance discriminability, there also are issues of fairness to consider (e.g., procedural justice, Tyler, 2003). Positive eyewitness identifications cannot be considered as providing independent evidence of guilt if the ID arises from the pressures of a lineup administrator (also see Hasel & Kassin, 2009). Unfortunately, little research has explored the effects of blind lineup administration on reliability, although Clark et al. (2013) reported that post-identification confidence was lower for those participants who were pushed or steered.
5.4 Lineup and Juror Instructions

5.4.1 Lineup Instructions

Numerous attempts have been made to improve eyewitness evidence through the use of instructions, with little success. Some of this research has focused on the instructions given to eyewitnesses, and other research has focused on the instructions given to jurors regarding how to weigh eyewitness evidence. We begin with biased versus unbiased instructions to eyewitnesses.

5.4.1.1 Biased vs. Unbiased Instructions

Biased instructions imply that the perpetrator is in the lineup, unbiased instructions do not (e.g., “The perpetrator may or may not be present”). Malpass and Devine (1981) conducted the first study comparing biased and unbiased instructions, and found a $d'$ advantage of 0.75 for unbiased instructions. Research that followed was in general agreement (e.g., Cutler, Penrod, & Martens, 1987; O’Rourke, Penrod, Cutler, & Stuve, 1989), which culminated in a meta-analysis in which Steblay (1997) concluded that unbiased instructions decreased choosing from target-absent lineups without decreasing correct IDs from target-present lineups. But more recent research has reached a different conclusion; a recent meta-analysis by Clark, Moreland, and Gronlund (2014) showed that the discriminability advantage for unbiased instructions aggregated across 23 studies was non-existent ($d' = -0.02$). In light of these results, it appears that biased versus unbiased instructions affect only response bias: Eyewitness are more conservative after receiving unbiased instructions.

Despite unbiased instructions not enhancing discriminability, the NAS report (National Research Council, 2014) nevertheless recommended that they be used:
“Witnesses should be instructed that the perpetrator may or may not be in the photo array or lineup...” (p. 73). To understand the rationale for this recommendation, one must consider the tradeoffs between the costs of errors (ID an innocent suspect, the failure to ID a guilty suspect) versus the benefits of correct decisions (ID the guilty suspect, reject a lineup that contains an innocent suspect). Most policymakers agree that it is more important to protect the innocent (limit IDs of innocent suspects) than it is to implicate the guilty. Blackstone (1769, p. 352) famously said that it is “…better that ten guilty persons escape than that one innocent suffer.” If the cost of a false ID is 10x greater than that of a miss, eyewitnesses should set a conservative criterion, and unbiased instructions should help accomplish that. However, it is important to point out that the choice of these utilities is a matter for society and policymakers, not for eyewitness researchers. Once the utilities are agreed upon, signal detection theory provides the machinery for converting the utilities, given the base rates of guilty and innocent suspects being placed into lineups, into an optimal criterion placement. For a recent review of these issues see Clark, Benjamin, Wixted, Mickes and Gronlund (2015).

5.4.1.2 Adding a “don’t know” option

Another instructional change that can be offered to eyewitnesses is the option of reporting that they “Don’t know.” Weber and Perfect (2012) conducted a study that examined the inclusion of a “Don’t know” option. Participants viewed a mock-crime video, a distracting video, and then a target-present or target-absent showup. Some participants were required to make a “yes” or “no” decision about the face in the showup, other participants were allowed to choose “don’t know” if they desired. Collapsed over retention interval (which was 3 minutes or 3 weeks), those participants with the option to choose “don’t know” made better discriminations ($d’ = .77$) than those participants...
without that option \(d' = .28\), response criterion position was more conservative, \(c = 0.75\) vs. \(c = 0.47\), respectively). But research needs to be conducted to examine the effects of an explicit Don't know option on reliability.

### 5.4.1.3 Appearance Change Instructions

Researchers also have explored the impact of instructing eyewitnesses that a perpetrator may have changed his appearance since committing a crime. Charman and Wells (2007) had participants view a video with four culprits; this was followed by four lineups (2 target-present and 2 target-absent, unbiased instructions). Half the participants received appearance change instructions and half did not. We computed \(d'\) and \(c\) from their data for each culprit, and then took the simple average across culprits. We found that the appearance change instructions made participants slightly more liberal, and slightly decreased discriminability. Molinaro, Arnsdorfer, and Charman (2013) conducted a similar study, and explicitly varied the amount of appearance change. They found that the appearance change instruction harmed discriminability and made responding more liberal, and that the effects were similar across targets that differed in the amount of appearance change. Finally, Porter, Moss, and Reisberg (2014) had participants view a mock crime, and used more extensive appearance change instructions. Experiment 2 included an explicit appearance change (addition of facial hair) for the target. Like the prior experiments, \(d'\) was lower for the appearance change condition (especially for the White culprit), although there was no response bias change. Overall, the data suggest that the appearance change instruction slightly decreases discriminability, and tends to make participants' response biases more liberal. Porter et al. reported that participants receiving the appearance change instructions were less confident in their choices; but an
assessment of the reliability of eyewitnesses receiving this instruction awaits further research.

5.4.2 Juror Instructions

Research has also been conducted on how jury instructions affect the evaluation of eyewitness evidence. Telfaire (United States v. Telfaire, 1972) instructions direct jurors to consider factors that could impact the accuracy of an eyewitness (e.g., viewing conditions), but fail to explain how these factors impact accuracy. Greene (1988) revised the Telfaire instructions, to simplify them, and added information about how various factors (i.e., the Biggers criteria) impact eyewitness accuracy. But compared to control participants that received no cautionary instructions, the Telfaire participants were no better at distinguishing between good and poor eyewitnesses. The revised instructions did, however, make mock jurors more skeptical overall. In a review of research evaluating the goal of improving the use of eyewitness evidence by jurors, Bornstein and Hamm (2012, p. 53) concluded “. . . the research on modifying instructions about witness identification has generally failed to accomplish this goal”.

New Jersey (2012) adopted an expanded, carefully designed (Schacter & Loftus, 2013) set of jury instructions (see http://www.judiciary.state.nj.us/pressrel/2012/jury_instruction.pdf). The goal of these instructions was to inform jurors about the current science of eyewitness memory, and how to use that knowledge to assess eyewitness testimony. Among other things, the instructions inform jurors that memory does not work like a video recording, and highlight the risks of making mistaken identifications. Papailiou, Yokum, and Robertson (2015) assessed the effectiveness of these instructions. Mock jurors viewed a 35-minute murder trial, which included either weak or strong eyewitness evidence. One-half of each group
was administered the New Jersey instructions, the other half was administered standard instructions. Jurors voted to convict about 25% of time given the standard instructions versus about 10% of the time given the enhanced instructions. More importantly, neither set of instructions aided mock jurors’ ability to distinguish between the weak and strong eyewitness evidence (despite clear differences in evidence quality, see Table 1 in Papailiou et al.).

In sum, instructions to eyewitnesses or jurors appear to make both more conservative. If the goal is to reduce false IDs of innocent suspects, this is commendable. But if the goal of ‘better’ instructions is to enhance the reliability of eyewitness evidence, it would appear that research efforts are better directed elsewhere. Moreover, if instructional variations are simply moving eyewitnesses’ response criteria around, an alternative perspective is to consider the response confidence reported by eyewitnesses. According to signal detection theory, changes in response confidence are a direct reflection of the underlying response criteria (Wixted et al., 2015). In other words, instead of conducting more research on the instructions administered to eyewitnesses, more research should be conducted on how best to assess eyewitness confidence, which, it appears, has the potential to significantly enhance the ability to distinguish between weak and strong eyewitness evidence if jurors can be made to rely on the initial, fairly assessed, identification evidence.

6 Conclusion

This chapter has covered a host of variables that affect eyewitness memory and introduced a new classification system for those variables. Going forward with research on eyewitness identification, it will be important to separately measure discriminability
(with ROC analysis) and reliability (with CAC analysis). Moreover, there is a need (see also Clark, 2008; Clark & Gronlund, 2015; Gronlund, Mickes, Wixted & Clark, 2015) to develop and test competing theories of discriminability and reliability, because eyewitness identification research generally has been guided by verbally specified theories or intuition, which makes it is difficult to extract definitive predictions (Bjork, 1973; Lewandowsky, 1993) and slows cumulative progress. A formally specified model, on the other hand, forces a theoretician to be explicit about a model’s assumptions, which makes predictions transparent and provides a check on reasoning biases (Hintzman, 1991).

Some progress has been made developing formally specified explanations. One such theory, WITNESS, was proposed by Clark (2003). WITNESS is a direct-access matching model (for an overview of this type of model, see Clark & Gronlund, 1996) with a signal-detection foundation. The model parameters are closely tied to the components relevant to eyewitness identification. The model has been used to explore filler selection, simultaneous and sequential lineups (Goodsell, Gronlund, & Carlson, 2010), and relative and absolute judgments (Clark, Erickson, & Breneman, 2011; Fife, Perry, & Gronlund, 2014). Wixted and Mickes (2014) extended a signal-detection model to the eyewitness domain. The theory proposed a diagnostic-feature-detection hypothesis to explain why discriminability is greater from simultaneous lineups. The idea is that by seeing all the lineup members at once, eyewitnesses can determine what features to pay attention to and what features are redundant across lineup members and therefore not diagnostic. For example, if all the individuals in the lineup are young African American males with shaved heads, putting attention on these shared (i.e., non-diagnostic) cues will not help an eyewitness attend to diagnostic cues that are unique to the perpetrator.
Many may find surprising what we have argued, that eyewitness identifications can be trustworthy, in the proper circumstances. The widely-held alternative view that eyewitnesses are always unreliable, arose because of the indisputable evidence regarding the malleability of memory, coupled with the large number of wrongful convictions due to faulty eyewitness IDs. But measurement errors involving how to assess discriminability (ROC analysis, not diagnosticity) and how to assess the relationship between confidence and accuracy (calibration or CAC analysis, not the point-biserial correlation) contributed to the fact that researchers reached some premature conclusions (for a review see Gronlund, Mickes, Wixted, & Clark, 2015). But the largest error, arguably, was to recommend discounting eyewitness confidence. Our review indicates that eyewitnesses often know to calibrate their confidence to reflect the likelihood that they are making an error, and that an identification made with high confidence is, in most circumstances examined to date, much more likely to be an identification of a perpetrator. Conversely, an eyewitness who makes an identification with low confidence is likely indicating that an identification may not be trustworthy. Although this chapter has not emphasized the low confidence end of the CAC plot, the implications of low confidence IDs also warrant careful consideration. Although a high confidence ID is likely to signal a guilty suspect, a low confidence ID should lead the police to consider that their suspect is innocent. By ignoring confidence, the criminal justice system has missed the opportunity to utilize information that can be of great value.

This is most painfully evident by considering the typical profile of a DNA exoneration case involving eyewitness misidentification, where the initial identifications often were made with low confidence (at best) that later inflated into a highly confident identification (see Garrett, 2011). By ignoring the initial level of confidence expressed on a
first, fair, test of memory, and instead presenting jurors with a highly confident witness in court, it is not surprising that the jurors in these cases reached the conclusion that the eyewitnesses were accurate in their misidentifications. In those cases, had initial confidence been made known, those wrongfully convicted individuals may not have suffered, and the actual perpetrators of these crimes could have been apprehended sooner.
7 References


Colloff, M., Wade, K., & Strange, D. (in press). Unfair lineups don’t just make witnesses more willing to choose the suspect, they also make them more likely to confuse innocent and guilty suspects. *Psychological Science*.


Predictions of face recognition are more accurate for members of our own race.

*Journal of Applied Research in Memory & Cognition, 1*, 158-162.


Johnson, K. J., & Fredrickson, B. L. (2005). “We all look the same to me” Positive emotions


Open Science Collaboration. (2015), Science, 349, aac4716. DOI: 10.1126/science.aac4716


Law, 34, 580-594.


doi:10.1016/0010-0285(90)90003-M


Table 1. Possible Decisions and Resulting Outcomes

<table>
<thead>
<tr>
<th>Eyewitness Decision</th>
<th>Identified the Suspect</th>
<th>Identified a Filler</th>
<th>Did Not Identify Anyone</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>True State</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Target Present</td>
<td>Correct ID (Hit)</td>
<td>Filler ID</td>
<td>Miss</td>
</tr>
<tr>
<td>Target Absent</td>
<td>False ID (False Alarm)</td>
<td>Filler ID</td>
<td>Correct Rejection</td>
</tr>
</tbody>
</table>


Figure 1. Variables that affect eyewitness accuracy categorized by stage of memory, and by which stage of the crime/criminal proceedings each occur. Depending on the point in the proceedings, results from different analyses are needed to guide decisions.

Note. ROC = receiver operating characteristic; CAC = confidence-accuracy characteristic
Figure 2. Six-person simultaneous lineup with five fillers and one suspect. If the suspect is the perpetrator, the lineup is target-present; if the suspect is innocent the lineup is target-absent.
Figure 3. Hypothetical ROC data from two procedures; Procedure A falls further from the line of chance performance (dashed line), therefore Procedure A has better discriminability than Procedure B. The solid lines represent the fit of a simple signal detection model (as shown in the top right panel of Figure 6).
Figure 4. (Top Panel) Detection (ROC) and detection-plus-identification (LROC) from Starr et al. (1975). (Bottom Panel) Detection (ROC) and detection-plus-identification (LROC) from Experiment 1 (collapsed across conditions) of Palmer et al. (2013). Figure from Wixted and Mickes (2015a, Figure 2).
Figure 5. Demonstration of objective discriminability of a showup and a lineup procedure. ROC curves of fair ("Innocent weak") data from Wetmore et al. (2015).
Figure 6. Simple signal detection based model. The means \((\mu_{\text{innocent}}, \mu_{\text{fillers}})\) and standard deviations of the filler/innocent suspect distributions are 0 and 1, respectively. The mean \((\mu_{\text{guilty}})\) and standard deviation of the target distribution are \(d\) and 1, respectively. If the memory strength of the most familiar lineup member exceeds the criterion \((c)\), an identification will be made. Unlike in a fair lineup, where fillers and innocent suspects are drawn from the same distribution (top panel on the left), in an unfair lineup, the fillers are drawn from a distribution that differs from the innocent suspect distribution (bottom panel on the left). In this case, \(\mu_{\text{fillers}} > \mu_{\text{innocent}}\) and results in a lower ROC for unfair lineups (bottom right panel).
Figure 7. Three ways to measure the positive predictive value of an identification: calibration analysis (top figure), confidence-accuracy characteristic (CAC) analysis (middle figure), and diagnosticity ratio (bottom figure).