Making Sense of Sequential Lineups: An Experimental and Theoretical Analysis of Position Effects

Brent M. Wilson, Kristin Donnelly, Nicholas J. S. Christenfeld, & John T. Wixted

University of California, San Diego

Author Note

Brent M. Wilson, Department of Psychology, University of California, San Diego. Kristin Donnelly, Department of Psychology, University of California, San Diego. Nicholas Christenfeld, Department of Psychology, University of California, San Diego. John T. Wixted, Department of Psychology, University of California, San Diego.

This work was supported by National Science Foundation grant SES-1456571 to John T. Wixted and by the Laura and John Arnold Foundation.

We thank the editor (Simon Farrell) and three reviewers (Jeff Starns, Adam Osth, and Yonatan Goshen-Gottstein) for their insightful comments on our paper.

Correspondence concerning this article should be addressed to Brent M. Wilson (b6wilson@ucsd.edu) or John T. Wixted (jwixted@ucsd.edu), Department of Psychology, University of California, San Diego, 9500 Gilman Dr., La Jolla, CA 92093

Abstract

As part of a criminal investigation, the police often administer a recognition memory task known as a photo lineup. A typical 6-person photo lineup consists of one suspect (who may or may not be guilty) and five physically similar foils (all known to be innocent). The photos can be shown simultaneously (i.e., all at once) or sequentially (i.e., one at a time). Approximately 30% of U.S. police departments have moved to using the sequential lineup procedure over the last 30 years, yet its theoretical underpinnings remain poorly understood. A simple signal detection model makes several unexpected predictions about how the sequential lineup procedure should affect the ability of eyewitnesses to discriminate innocent from guilty suspects. For example, empirical discriminability (area under the receiver operating characteristic) should decrease as the position of the suspect in the lineup increases. In addition, under some conditions, a fair sequential lineup should not yield higher discriminability than a single-person (non-lineup) recognition test known as a showup. The results of two experiments reported here confirmed these predictions. Counterintuitively, even though empirical discriminability decreased as the suspect's sequential position increased, a signal detection model fit to the data indicated that theoretical discriminability exhibited a small effect in the opposite direction (increasing with the sequential position of the suspect). The latter result is consistent with a diagnostic feature-detection theory of eyewitness identification.

Keywords: Recognition Memory; Signal Detection Theory; ROC Analysis; Diagnostic Feature-Detection Theory; Eyewitness Memory

Making Sense of Sequential Lineups: An Experimental and Theoretical Analysis of Position Effects

Nowadays, the most common eyewitness identification procedure in the United States is a photo lineup, which has largely replaced the live lineups the police once used. A photo lineup consists of a picture of one suspect (the person who the police believe may have committed the crime) plus several additional photos of physically similar people who are known to be innocent. The photos can be shown all at once – the traditional simultaneous lineup, developed by the police long ago – or one at a time – the newer sequential lineup, developed by experimental psychologists in 1985 (Lindsay & Wells, 1985). In lab-based studies of the sequential lineup, a stopping rule is typically used such that the first photo that is identified terminates the procedure.

Researchers usually evaluate competing lineup formats using mock-crime laboratory experiments in which participants witness a staged crime and are later shown a photo lineup in which the perpetrator is either present or absent. A target-present lineup includes the perpetrator along with (usually 5) similar foils; a target-absent lineup is the same except that the perpetrator is replaced by another similar foil who serves as the designated innocent suspect. For decades, diagnostic accuracy was assessed using a statistic known as the diagnosticity ratio (DR), which is the hit rate (HR) divided by the false alarm rate (FAR). The HR is the proportion of target-present lineups that resulted in a correct identification of the guilty suspect. For example, if 70% of target-present lineups resulted in a correct ID of the guilty suspect, 20% resulted in an incorrect ID of a foil, and 10% resulted in no ID, the HR would be .70. The FAR is the proportion of target-absent lineups that resulted in an incorrect identification of the innocent suspect. For example, if 6% of target-absent lineups resulted in an incorrect ID of the innocent suspect, 30% resulted in an incorrect ID of a foil, and 64% resulted in no ID, the FAR would be .06.[1] In the seminal study on this issue, Lindsay and Wells (1985) reported that for sequential lineup, HR = .50 and FAR = .17 (DR_{SEQ} = .50 / .17 = 2.94), whereas for the simultaneous lineup, HR = .57 and FAR = .42 (DR_{SIM} = .57 / .42 = 1.36). The higher DR for sequential lineups is the basis for what came to known as the "sequential superiority effect" (e.g., Steblay, Dysart, & Wells 2011).

In recent years, it has become widely appreciated that the DR conflates discriminability (the ability of an eyewitness to discriminate innocent from guilty suspects) with response bias (the overall tendency to identify anyone from a lineup) (Gronlund, Wixted, & Mickes, 2014). Because it increases as responding becomes more conservative, a higher DR does not provide evidence that one procedure is superior to another. Using signal detection theory as a guide, researchers are now more likely to compare the diagnostic accuracy of competing lineup procedures by measuring the area under the receiver operating characteristic (ROC). The signal detection approach has also brought to light some unusual properties of the sequential lineup procedure that had previously gone unnoticed (Rotello & Chen, 2016). We pursue those unusual properties in some detail here, but we first briefly consider how signal detection theory is ordinarily applied to a (non-lineup) recognition memory task.

Signal Detection Theory of Recognition Memory

Signal detection theory has been applied to old/new recognition memory since Egan's (1958) seminal report more than half a century ago. A standard old/new memory test consists of targets (previously presented list items) randomly intermixed with foils (novel items), each presented for an individual "old/new" decision. Theoretically, the memory signal for a target is

[1] If the innocent suspect is another randomly selected foil, the FAR can also be estimated by counting all identifications from target-absent lineups and then dividing by lineup size. In this example, the FAR would be (.06 + .30) / 6 = .06.

drawn from a Gaussian distribution with a relatively high mean, whereas the memory signal for the foil is drawn from a Gaussian distribution with a relatively low mean (Figure 1). The decision to either identify or reject the test item is made in relation to a decision criterion (*c*) placed somewhere on the memory strength axis. At one extreme, with the criterion placed far to the left (very liberal response bias), a participant would almost always respond “old,” in which case both the hit rate and the false alarm rate would be ~1. That is, all targets would be correctly classified as old, but all foils would be incorrectly classified as old. At the other extreme, with the criterion placed far to the right (very conservative response bias), a participant would almost always respond “new,” in which case both the hit rate and the false alarm rate would be ~0. That is, all targets would be incorrectly classified as new, and all foils would be correctly classified as new.

When conducted in the context of eyewitness identification, an old/new recognition test is called a showup. In a photo showup, the test item is either the person who committed the crime (a photo of the target) or an innocent suspect (a photo of a foil). In the eyewitness context, an “old” decision is an identification (ID), and a “new” decision is a non-identification (non-ID). The main difference between a standard old/new recognition experiment and a showup is that in the former, each subject studies and is tested with many old and new items, whereas in the latter, each study item is viewed by many subjects who are then tested with a single old or new item. Thus, item variance is a significant factor in a typical list-memory study, whereas subject variance is a significant factor in a showup experiment. Despite these differing sources of variance, the signal detection conceptualization of the problem is the same in the two cases. As with the standard old/new recognition memory test, a liberal response bias (with *c* set far to the left) would yield hit and false alarm rates close to 1, and a conservative response bias (with *c* far to the right) would yield hit and false alarm rates close to 0.

A Signal Detection Model for Lineups

The standard signal detection-based interpretation of lineup performance involves additional considerations because of the presence of foils on a given test trial. Thus far, this model has most often been applied to simultaneous lineups. In a 6-person simultaneous lineup, all six faces are presented together. The simplest signal detection model of decision making on this task is still grounded in the basic model shown in Figure 1. The lineup version of the model holds that, for a target-present lineup (consisting of one guilty suspect plus five foils), the memory signal for the suspect is randomly drawn from the target distribution, and the memory signals for the five foils are randomly drawn from the foil distribution. For a target-absent lineup (consisting of one innocent suspect plus five foils), the memory signals for all six lineup members are randomly drawn from the foil distribution. The reason is that for a fair target-absent lineup, from the witness's perspective, the innocent suspect is effectively a foil (i.e., the innocent suspect – like the foils – is someone who physically resembles the perpetrator but did not commit the crime). Using the simplest decision rule, the most familiar face in the lineup – that is, the face that generates the strongest memory signal (i.e., the MAX signal) – is identified if it exceeds the decision criterion (*c*). This simple model is known as the Independent Observations model (Duncan, 2006; Macmillan & Creelman, 2005; Wixted, Vul, Mickes, & Wilson, 2018).

A hit – also known as a correct ID – occurs when the MAX signal in a target-present lineup is generated by the guilty suspect and exceeds the decision criterion. A false alarm – also known as a false ID – occurs when the MAX signal in a target-absent lineup is generated by the innocent suspect and exceeds the decision criterion. A foil ID for either lineup occurs if one of the foils happens to generate the MAX memory signal in the lineup and the strength of that signal exceeds the criterion. If none of the faces in the lineup generate a memory signal that exceeds the criterion, no ID is made (i.e., the lineup is rejected).

Using a simultaneous lineup, if eyewitnesses have a conservative decision criterion, both the hit rate and the false alarm rate will be low (and will reach 0 in the limit), as will the foil ID rate. At the other extreme, however, an infinitely liberal criterion will yield higher hit and false alarm rates but will not yield hit and false alarm rates of 1.0, as would be the case for a showup. So long as the identity of the suspect is not highlighted (i.e., the lineup is fair and the administrator does not provide any clue as to which photo is the suspect), the hit rate would be unlikely to reach 1.0 because the liberal eyewitness must determine which of the six faces in the target-present lineup is the perpetrator. Thus, a liberal eyewitness who failed to form a strong memory of the perpetrator would stand a good chance of landing on a filler. The false alarm rate would also fall well below 1.0. In a fair target-absent lineup, the maximum false alarm rate under extremely liberal responding would be only 1 / 6 = .167. Despite these upper-bound constraints, it is still true that the hit rate and the false alarm rate will monotonically increase as eyewitnesses move from more conservative to more liberal responding. The hit and false alarm rates generated by varying the decision criterion, when plotted against each other, trace out the lineup ROC curve.

ROC data, including lineup ROC data, can be generated in more than one way, such as using different instructions across different conditions to induce liberal, neutral, or conservative response biases. For example, a liberal response bias could be induced by instructing participants to choose the perpetrator even if they have to guess; a neutral response bias could be induced by simply indicating that the perpetrator may or may not be in the lineup; and a conservative response bias could be induced by instructing participants not to make an identification unless they were highly confident the person they were selecting was actually the perpetrator. These conditions would yield high, medium, or low hit and false alarm rates, respectively, and plotting the hit rate vs. the false alarm rate would yield a 3-point ROC. This type of ROC is often called a "binary" ROC (e.g., Dube & Rotello, 2012) because the recognition decisions within a given condition are binary (ID vs. no ID). Alternatively, and more commonly, confidence ratings can be used to generate the ROC data from a single condition. Consider a 3-point confidence scale for positive IDs of the suspect made with high, medium, or low confidence. The conservative (leftmost) ROC point would be obtained by counting only IDs made with high confidence; the middle ROC point would be obtained by counting IDs made with medium *or* high confidence, and the liberal (rightmost) ROC point would be obtained by counting all IDs (made with high, medium, or low confidence).

For decades, research from the basic memory and perception literatures has found that various strategies for generating ROC data tend to yield the same (or at least similar) curves (e.g., Benjamin, Tullis & Lee, 2013; Dube & Rotello, 2012; Koen & Yonelinas, 2011; Swets, Tanner & Birdsall, 1961). Recently, Mickes et al. (2017) demonstrated that the same[2] is true for simultaneous lineup ROC data generated using instructions or confidence ratings (i.e., the binary ROC was similar to the ratings-based ROC). Figure 2 reproduces their results, a ratings ROC from confidence ratings (Figure 2A) and a binary ROC from an instructional biasing manipulation with 4 conditions (Figure 2B). The smooth curve drawn through the data is the same for both plots. Although there is some evidence for a reduction in discriminability for the

[2] Mickes et al. (2017) found that discriminability was somewhat reduced when biasing instructions were intended to make people extremely liberal or conservative, theoretically due to increased criterion variability across participants.

extreme biasing conditions in Figure 2B, the ROC paths traced out using these two methods are similar. The same is definitely not true for sequential lineups.

A Signal Detection Analysis of Sequential Lineups

The seemingly simple change from presenting photos simultaneously to presenting them sequentially (and with a stopping rule) raises some surprisingly complicated theoretical and empirical issues. Nevertheless, it is worth thinking through those complexities because 30% of U.S. law enforcement agencies have now adopted the sequential procedure (Police Executive Research Forum, 2013). Given the applied significance of this recognition memory procedure, it seems important to understand how sequential lineups affect underlying (theoretical) discriminability (*d'*) and, separately, empirical discriminability (i.e., area under the ROC, or AUC). Although *d'* and AUC ordinarily agree about the effect of an independent variable on discriminability, they are capable of reaching opposite conclusions (Wixted & Mickes, 2018). As shown later, the sequential procedure can yield that unusual outcome.

Prior Signal Detection Analyses of the Sequential Procedure. There is already reason to believe that much of what we commonly understand about ROC analyses of recognition memory breaks down with a sequential lineup. Using a signal detection model, Rotello and Chen (2016) simulated ROC data generated by the sequential procedure. As noted above, the simplest version of the model for simultaneous lineups relies on a MAX decision rule according to which the face in the lineup that generates the strongest memory signal is identified if the strength of that signal exceeds a decision criterion. For the sequential procedure, by contrast, a first-above-criterion decision rule is used because of its stopping rule. That is, the first face that generates a memory signal strong enough to exceed the criterion is identified, terminating the procedure.

Rotello and Chen (2016) showed that the constraints in a sequential lineup introduced by the stopping rule are such that changing the overall decision criterion from liberal to conservative across conditions (i.e., generating binary ROC data) traces out an unusual curve. This curve does not increase monotonically (as the one in Figure 2B does) but instead increases at first and then decreases towards the diagonal line of chance performance. What explains that pattern? Consider, for example, a group of eyewitnesses instructed to use an infinitely conservative criterion. These eyewitnesses never make an identification so the hit rate and false alarm rate would both be 0, as it would be for any signal detection task. Next consider the opposite extreme, namely, a group of eyewitnesses instructed to use an infinitely liberal decision criterion. Because these witnesses are always going to make an identification regardless of the strength of the memory signal generated by a test item, they will always identify the first face in the lineup, thereby terminating the procedure at that point. For a subset of these infinitely liberal witnesses, the suspect will have been randomly assigned to appear in position 1. A guilty suspect in position 1 will always be correctly identified by this subset of liberal eyewitnesses. Similarly, an innocent suspect in position 1 will always be incorrectly identified.

Of course, the suspect will not always appear in the first position. For other witnesses, the suspect will appear somewhere in positions 2 through 6. When the suspect is placed in those positions, the stopping rule is such that infinitely liberal eyewitnesses will have no opportunity to make any additional identifications, having already identified a filler in position 1. If we assume that there is an equal probability of the suspect appearing in any of the six positions, this means that a group of eyewitnesses with an infinitely liberal criterion will have hit and false alarm rates of 1/6 ≈ .167. Thus, the instruction-based binary ROC for a fair sequential lineup ranges from [0, 0] when an infinitely conservative criterion is used to [.167, .167] when an infinitely liberal criterion is used.

As shown by Rotello and Chen (2016), for intermediate settings of the criterion, the ROC will bow up and away from the diagonal line of chance performance: the sequential binary ROC curve is non-monotonic. Figure 3 reproduces results shown in Figure 6 of Rotello and Chen (2016). The figure shows the predicted ROC curves for simultaneous and sequential lineups using an equal-variance model with *d'* = 1.5 as the criterion is swept over the full range from liberal to conservative. Note that despite the equivalence in underlying *d'*, the area under the ROC would be quite different for the two procedures (Wixted & Mickes, 2018), clearly favoring the simultaneous procedure. Thus, even if *d'* were the same for the two procedures, then, according to this model, it would be a mistake for the police to believe that it would make sense to adopt the sequential procedure. For any given false alarm rate, the simultaneous procedure would be able to achieve the same or higher hit rate. Critically, this reduction in area under the ROC is not due to reduced psychological (i.e., underlying) discriminability; it arises entirely from the physical constraint imposed by the standard first-identification-only stopping rule.

The unusual sequential ROC simulated by Rotello and Chen (2016) sets the stage for the detailed theoretical and empirical analysis of the sequential procedure that we report here. To guide our inquiry into these issues, we also rely on a simple signal detection model using a first-above-criterion decision rule to generate predictions about sequential performance. Later, we report the results of two large-*N* experiments to test whether the predictions made by Rotello and Chen (2016) and the additional predictions we generate next are accurate. The parameters used in our simulations are set to approximate the empirical data we consider later. To begin, following Rotello and Chen, we use a signal detection model of binary ROC data in which a different decision criterion is used in each of several different conditions (e.g., as could be achieved using different instructions across conditions).

Figure 4A illustrates this simple equal variance[3] signal detection model, with $d' = 1.74$, and with the decision criteria set to be liberal ($c = -0.25$), neutral ($c = 0.93$), or conservative ($c = 1.75$). As we earlier assumed for the simultaneous procedure, the model assumes that for each face in a target-present lineup, one value is randomly drawn from the target distribution (the guilty suspect) and five values are randomly drawn from the foil distribution, and for each face in a target-absent lineup, six values are randomly drawn from the foil distribution. However, in contrast to a simultaneous lineup, the six values for a given sequential lineup are individually compared against a preset criterion in the order in which they were randomly sampled. If one of these memory-match values exceeds the criterion, an ID is made and the test is terminated (i.e., a first-above-criterion decision rule was used). Hit and false alarm rates are computed in the manner described earlier for the simultaneous procedure. That is, the hit rate is the proportion of all target-present lineups in which eyewitnesses identified the guilty suspect, and the false alarm rate is the proportion of all target-absent lineups in which they identified an innocent suspect. For the time being, we ignore foil IDs, though they are always taken into consideration when fitting signal detection models to lineup data (simultaneous or sequential). Because this simulation was performed in the same manner as the one performed by Rotello and Chen (2016), it is not surprising that it yielded a non-monotonic ROC similar to the one they reported. Figure 4B shows the predicted binary ROC points, and it is clear that it traces out a non-monotonic path.

[3] In analyzing our data, we will allow for an unequal variance model in order to better fit the data. However, this change does not substantively affect any of our model predictions. Although there is nothing special about the particular *c* values used, we chose them because they result in approximately the same overall false alarm rates that were observed in our empirical data presented later.

However, as we explain next, there is more to the surprising story of how the stopping rule affects ROC data generated by the sequential procedure.

One Binary ROC Yields Multiple Confidence-based ROCs. Whereas Figure 4B reproduces what has been shown before, Figure 4C shows another unusual feature of sequential ROC data. More specifically, Figure 4C shows the confidence-based ratings ROC associated with each instruction-based binary ROC. The rightmost point of each rating's ROC in Figure 4C is the same as the corresponding point on the binary ROC in Figure 4B. Consider, for example, subjects assigned to the liberal instructional biasing condition (leftmost criterion in Figure 4A). Their overall hit and false alarm rates would correspond to the rightmost ROC point in Figure 4B, but if those subjects were asked to provide confidence ratings, their confidence-based ROC would correspond to the lowest curve in Figure 4C. For subjects assigned to the conservative instructional biasing condition (rightmost criterion in Figure 4A), their overall hit and false alarm rates would correspond to the leftmost ROC point in Figure 4B. However, if they were asked to provide confidence ratings, their confidence-based ROC would correspond to the highest ROC in Figure 4C. In other words, each between-condition, binary ROC point has its own unique confidence-based ROC.

If we performed this exact same simulation for the simultaneous lineup wherein the overall criterion is either liberal, neutral, or conservative, all three confidence-based ROCs would fall on top of one another. In other words it would look like a single ROC curve as it does for the simultaneous lineup in Figure 3. However, unlike all other cases that we are aware of, for the sequential procedure, a singular binary ROC like that shown in Figure 4B does not have a corresponding singular confidence-based ROC. Instead, there is one confidence-based ROC for each binary ROC point, as depicted in Figure 4C, where the dotted lines correspond to the binary

ROC in Figure 4B. So far as we know, this dissociation between confidence-based and binary ROC data is unique to the sequential procedure.

Although we illustrated this issue with respect to hypothetical instruction-based binary ROCs vs. confidence-based ratings ROCs, the same issue can be illustrated using data from the neutral response biasing condition only. We do so next because it corresponds directly to how we later analyze empirical data obtained using the sequential procedure. Figure 5 shows a confidence scale ranging from -100 (sure the face was not seen) to +100 (sure this is the face of the perpetrator). In a sequential lineup procedure, the witness can be asked to provide a confidence rating using this scale for each face in the lineup. The standard stopping rule – but not the only possible stopping rule – stipulates that any positive ID associated with a confidence rating greater than 0 terminates the procedure (such that any IDs of subsequent faces in the lineup would not count). The confidence-based ROC for positive suspect IDs in the neutral condition would correspond to the middle ROC curve shown in Figure 4C.

Given the continuous nature of the confidence scale shown in Figure 5, it should be clear that defining the stopping rule to consist of any ID greater than 0 is an arbitrary decision. Indeed, using that same set of data from the neutral condition, one could specify a different decision criterion for counting positive IDs in separate analyses, with each analysis corresponding to a different stopping rule. For example, a conservative criterion could be specified by counting only IDs made with a confidence rating of 80 or more. Moving the criterion from 0 to a more conservative setting of 80 for either showups or simultaneous lineups would simply move the ROC point to the left on the same ROC (i.e., the ROC curve itself would not change). In fact, this is precisely how a conservative ROC point would be generated for a confidence-based ROC. However, given the nature of the stopping rule, setting a more conservative criterion for the sequential lineup has additional consequences: IDs made in an earlier sequential position with a confidence rating of less than 80 would no longer cancel subsequent IDs (because those earlier IDs would now be treated as effective non-IDs). The resulting confidence-based ROC for this conservative decision rule would cover a narrower range (i.e., the maximum hit and false alarm rate would be lower) and would also be elevated compared to the neutral ROC curve, as shown in Figure 4C. Similarly, setting the decision criterion to -80 (a very liberal position) has the opposite effect, lowering and extending the confidence-based ROC to the right.

In the empirical analyses we present later, we analyze the data in this manner. That is, we analyze data from a sequential ROC study that used neutral instructions and collected confidence ratings using a scale like that shown in Figure 5. In real-world scenarios, the police could opt to establish binary decision criteria in the same way that we did so long as they collected confidence for each ID. For example, neutral Jurisdiction A (the typical case) might use the most intuitive rule and count any ID made with a confidence rating greater than 0. By contrast, conservative Jurisdiction B might decide not to count an ID unless it was made with a confidence rating greater than 80 (in hopes of reducing misidentifications of the innocent), whereas liberal Jurisdiction C might decide to count any ID made with a confidence rating greater than -80 (in hopes of increasing correct identifications of the guilty). If they could be measured, the hit and false alarm rates from these three jurisdictions would create a 3-point binary ROC. Yet each binary ROC point would have its own separate confidence-based ROC.

Sequential Position Effects. The story of confidence-based ROCs generated using the sequential procedure becomes even more complicated (and more surprising) when we consider the ROC curves separately by suspect position (i.e., where the suspect appears in the sequence of six faces). For the simulated analyses presented thus far, suspects were randomly assigned to position and the data were aggregated across position. This corresponds to how sequential ROC data have been plotted in previous empirical research (e.g., Mickes et al., 2012). However, the simple signal detection model can also predict ROC curves separately for trials in which the suspect (innocent or guilty) appears in position 1, position 2, position 3, and so on.

We used the first-above-criterion signal detection model in a simulation to investigate position effects by first assigning the suspect (either innocent or guilty) to a randomly selected position (1 through 6) and then determining the probability that the suspect would be identified, separately for each position. If, for example, the suspect was randomly assigned to position 3, the probability of a suspect ID would be the probability that a foil memory strength did not exceed the neutral decision criterion in either position 1 or position 2 times the probability that the suspect memory strength in position 3 did exceed the neutral decision criterion. If no filler IDs happen to occur in the first two positions and a suspect ID is made in position 3, the corresponding confidence rating is determined by the highest confidence criterion exceeded by the memory strength of the suspect. Keep in mind that, sometimes, the suspect memory strength will far exceed the decision criterion (and will be the MAX face in the lineup) but will not be identified because a preceding filler memory strength also exceeded the criterion, even if only slightly. This is the constraint imposed by the stopping rule.

Figure 6A shows the six separate position-specific ROC curves for the neutral response bias condition, as predicted (via simulation) by the first-above-criterion signal detection model. As suspect position increases from 1 to 6, each successive rightmost ROC point – that is, each successive overall hit and false alarm rate – falls below the ROC for the immediately preceding position (*not* on the same ROC), and also moves to the left. Thus, for a given false alarm rate (e.g., false alarm rate = .05), the area under the ROC decreases with increasing suspect position.

As noted above, the overall hit and false alarm rates both decrease as suspect position increases because foil IDs that occur in earlier positions remove opportunities to make suspect IDs in later positions. The position-specific ROC data would not seem at all unusual if the reduction in the overall hit and false alarm rate with increasing sequential position were such that the resulting ROC point fell on a more conservative location on the ROC for the preceding position. In that case, the effect of foil IDs made in earlier positions (canceling opportunities to make suspect IDs in later positions) would be the effective equivalent of adopting a more conservative decision criterion. However, as the innocent and guilty suspects are placed later in the lineup, the overall correct and false alarm rates instead drop to a more conservative location on a *lower* ROC. To remain on the same ROC but at a more conservative location, the false alarm rate would have to decrease to a proportionately greater extent than the hit rate decreased. This is just another way of saying that the diagnosticity ratio (hit rate / false alarm rate) increases as responding becomes more conservative while holding discriminability constant (i.e., while staying on the same ROC). Indeed, that property of the DR – namely, that it increases with more conservative responding while holding discriminability constant – is why it was a mistake for the field to once rely on that measure to proclaim a sequential superiority effect. However, as the position of the suspect in the sequential lineup increases, the stopping rule does not reduce hit and false alarm rates in the same way that using a more conservative decision criterion does. Instead, foil IDs that occur in earlier sequential positions cancel opportunities to make later suspect IDs *to an equal extent* for innocent and guilty suspects. For example, if the innocent and guilty suspects are in position 2, and if 25% of witnesses ID a foil in position 1 for both target-present and target-absent lineups (which, at position 1, are identical), then, all else being equal, both the hit rate and the false alarm rate will decrease by 25% compared to position 1. Table 1 shows the overall hit and false alarm rates (i.e., the rightmost ROC points) and their corresponding diagnosticity ratios for the simulated ROC curves shown in Figure 6A. Although the hit and false alarm rates drop dramatically with increasing position (because opportunities to make a suspect ID steadily decrease), the diagnosticity ratio for each rightmost ROC point remains roughly constant at about 4.5, which further illustrates why the DR is not an appropriate measure of discriminability. Empirical discriminability is clearly decreasing with suspect position in these simulated data (despite underlying *d'* remaining constant), yet the DR misses that fact.

As a practical matter, the theoretical data in Figure 6A suggest that if the police always placed the suspect in position 6, for example, they would achieve what appears to be a desirable outcome (namely, a low false alarm rate, protecting the innocent), but the cost would be a dramatic reduction in empirical discriminability. Indeed, one could achieve both a lower false alarm rate and a much higher hit rate by instead placing the suspect in position 1 and using a more conservative location on the confidence scale as the criterion for counting an ID. For example, in Figure 6A, the fifth point from the right on the position 1 ROC has a hit rate of .56 and a false alarm rate of just under .06. This has both a higher hit rate and a lower false alarm rate than the overall hit rate of .30 and false alarm rate of .07 when the suspect is instead placed in position 6.

The Interaction of Response Bias and Position Effects. Earlier, we replicated simulations performed by Rotello and Chen (2016), showing that according to a simple signal detection model, for a given *d'*, the binary ROC generated by manipulating response bias is non-monotonic (Figure 4B). In addition, we just showed that according to the same model, the confidence-based ROC changes as a function of sequential position. Thus, according to this model, both the overall decision criterion and, separately, the sequential position of the suspect within the lineup affect the empirical ROC. We now show that these two effects (the effect of the placement of the overall decision criterion and, separately, the effect of suspect position) theoretically interact with each other.

As shown in Figure 6B, sequential position effects become even more exaggerated when a liberal criterion is used (the criterion placed at -80 in Figure 5). If eyewitnesses use a liberal criterion, it is very likely that one of the fillers in the first or second position will exceed the decision criterion. This makes it almost impossible for witnesses to ever have the opportunity to identify a face presented later in the lineup. By contrast, Figure 6C shows that position effects are minimized when a conservative criterion is used (the criterion placed at 80 in Figure 5). This is because foil identifications are much less likely to occur, which allows witnesses to have the opportunity to identify suspects appearing later in the lineup.

Note that the number of confidence-based ROC points decreases as the overall decision criterion becomes more conservative. For example, when placed at a very conservative position (80), only 2 confidence-based ROC points – namely, IDs made with confidence greater than 80 and IDs made with confidence greater than 90 – can be computed. When placed at a very liberal position (-80), as many as 18 confidence-based ROC points can be computed.

As can be observed in Figure 6A-6C, empirical discriminability is highest when the innocent and guilty suspects appear in the first position of the lineup and lowest when they appear in the last position (though this effect is minimized for the conservative condition). When the suspect appears in the first position, the sequential lineup is theoretically equivalent to a showup. In fact, if the police always placed the suspect in position 1, the procedure would presumably differ from a showup only in that witnesses might adopt a more conservative decision criterion, knowing that additional photos will be shown if no ID is made for the first photo. If the police instead always placed the suspect in a later position, the basic signal detection model predicts that empirical discriminability should be *reduced* relative to a showup. The predicted reduction in discriminability is due entirely to the structural constraints of the first-identification-only stopping rule and does not imply that psychological discriminability (i.e., *d'*, the ability to tell the difference between innocent and guilty suspects) necessarily changes as the lineup progresses. The question of how psychological discriminability changes (or not) as a function of sequential position is an independent issue, and we turn to a consideration of that issue next.

The Effect of Sequential Position on Psychological Discriminability (*d'*). Previous research suggests that, even in the absence of a structural constraint, psychological discriminability might change as a function of sequential testing position. For example, studies using standard list-learning paradigms have found that recognition memory performance tends to decline as a function of test trials (Peixotto, 1947; Murdock & Anderson, 1975; Criss, Malmberg, & Shiffrin, 2011). Osth, Jansson, Dennis, and Heathcote (2018) recently explored this phenomenon and found that it was primarily due to prior test items causing the contextual representation that cues memory to drift. Declining performance over the course of testing might also occur because of test-item interference and/or changes in speed-accuracy thresholds for later test trials. Regardless of which explanation applies, this line of research suggests that underlying discriminability might, if anything, decrease as sequential position of the suspect in a lineup increases. If so, sequential lineups would be worse than showups because both psychological factors and structural constraints would contribute to lower performance for later positions.

Another theory predicts that psychological discriminability might increase as a function of sequential position. Any such improvement in psychological discriminability would serve as a countervailing force against the structural constraints that lower performance for later positions. Diagnostic-feature-detection theory (Wixted & Mickes, 2014) holds that seeing faces that match the general description of the suspect teaches eyewitnesses which features are unlikely to be helpful for making an identification –namely, the features that are common to everyone in the lineup. These common features are the ones that were central to constructing the lineup in the first place (i.e., the selection of physically similar foils who all correspond to the description of the perpetrator). Precisely because these features are shared, taking them into consideration would reduce the ability of witnesses to distinguish between innocent and guilty suspects, whereas discounting those features would have the opposite effect (enhancing discriminability).

Diagnostic feature-detection theory was advanced to explain why simultaneous lineups typically yield a higher ROC than sequential lineups, even when responding is conservative (such that constraints imposed by the stopping rule are minimized), as it was in Mickes et al. (2012). The basic idea is that when faces are presented simultaneously, it is easier to detect (and then discount) non-diagnostic facial features than when faces are presented sequentially. When the suspect appears in the first position, there is no possibility of learning what features are diagnostic, but as the sequential lineup unfolds, the theory predicts that the same phenomenon should eventually emerge. Thus, if one assumes that faces need not be presented simultaneously for participants to notice and discount non-diagnostic features, the prediction would be that psychological *d'* should increase with increasing sequential position, an effect that could be easily masked by the large constraints on the empirical ROC imposed by the stopping rule.

Previous research on how memory performance changes as a function of sequential position has not distinguished between empirical and psychological discriminability. Doing so requires both computing area under the ROC to measure empirical discriminability and fitting a model to the ROC data to measure psychological discriminability. Instead, only measures of empirical performance – the diagnosticity ratio originally and, later, partial area under the ROC – have been used. Moreover, the results have been somewhat inconsistent. Carlson, Gronlund, and Clark (2008) found that the diagnosticity ratio increased for later positions of the sequential lineup, which is consistent with increasing discriminability but might instead simply reflect conservative responding as a function of lineup position. For example, a similar explanation in terms of conservative responding has been proposed to explain the effect of providing a verbal description of a face (Clare & Lewandowsky, 2004). Theoretically, providing a verbal description allows people to realize that the task is challenging, which causes them to behave more cautiously in their subsequent responding. Critically, this more conservative responding can occur whether or not discriminability is affected (Wilson, Seale-Carlisle, & Mickes, 2017).

In a study that reported more convincing evidence of suspect position on discriminability, Gronlund et al. (2012) performed ROC analysis separately for when the suspect appeared in either the second or fifth position of a sequential lineup. They found that empirical discriminability (measured as partial area under the ROC curve) was significantly higher for position five than position two, as predicted by the diagnostic feature-detection theory. Given the constraints imposed by the stopping rule (which, if anything, create a force in the opposite direction), this result indicates that underlying discriminability must have increased for later positions. However, a position effect on empirical discriminability measured by partial area under the ROC was not observed in other studies of the sequential procedure (Carlson &

Carlson, 2014; Dodson & Dobolyi, 2013). Thus, the empirical picture is mixed. Given that there is some theoretical reason to believe both that *d'* should increase with sequential position (due to diagnostic feature detection) and that area under the curve should decrease with sequential position (because of structural constraints imposed by the stopping rule), a mixed empirical literature may be unsurprising. What *is* somewhat surprising is that evidence of a decreasing area under the curve with increasing sequential position (Figure 6A) has never been empirically observed. However, we observed it in the study described next. Moreover, the very same data provided evidence that despite the decreasing area under the curve with increasing sequential position, *d'* changed in the opposite direction.

Experiment 1

The first experiment was a large-*N* investigation of eyewitness identification performance using a sequential lineup. Many participants were tested so that the data could be productively analyzed separately by sequential position. Each participant first watched a simulated crime video and then attempted to identify the perpetrator from a 6-member sequential lineup. All participants were presented with (and supplied confidence ratings for) all 6 faces in the lineup, which allowed us to specify different stopping rules. We created different points on the "binary" ROC by setting a different confidence criterion for counting an ID (namely, IDs with a confidence rating > -80, > 0, or > 80), as illustrated in Figure 5.

Method

Participants

A total of 7,174 subjects were tested using MTurk, but some answered the attention-check question incorrectly and were excluded, leaving a total of 6,530 participants. Of those, 3,258 were presented with a target-present lineup and 3,272 with a target-absent lineup.

Participants were each paid $0.25 for completing the task.

Materials and Procedure

All participants watched an eight-second video of a woman spray-painting graffiti. Participants then completed a distractor task (unscrambling the names of ten U.S. states) for approximately 2 minutes before beginning the sequential lineup test. Just prior to the lineup, participants received neutral identification instructions (indicating that the target from the video may or may not be in the lineup) and were also instructed on how to use the confidence scale. Participants were randomly assigned to only one lineup test – either a *target-present* sequential lineup or a *target-absent* sequential lineup. The target-present lineups contained a photo of the suspect (the woman in the video) and five filler photos, which were randomly selected from a set of 113 description-matched photos. The position in which the suspect appeared was randomized. *Target-absent* lineups consisted of six filler photos, again randomly selected from the set of 113. All participants were shown each photo one at a time. For each face, they were asked whether or not it was the woman from the video, and to indicate their level of confidence using a scale from -100 to 100 (Figure 5). Note that participants were asked to provide a confidence rating for all 6 faces, which means that the procedure was not actually terminated if a face was identified early in the sequence. As is typically true, participants were not told how many faces they could expect to see, and they were not told that only their first ID would count.

Results

Initially, we simply plotted the ROC data aggregated over sequential position. Later, we plot the data separately by position and also analyze the data by fitting a signal detection model. We begin by plotting the binary ROC, generating different binary ROC points by setting a different decision criterion on the confidence scale. As a reminder, if a simultaneous lineup or a showup were used, creating ROC data using this approach (instead of using instructions to manipulate the criterion across conditions, for example) would simply trace out the confidence-based ROC. That is, in fact, how the confidence-based ROC is typically created. However, because the sequential procedure uses a stopping rule, the binary analysis (i.e., using confidence ratings to set the value above which an ID is counted but otherwise ignoring differences in confidence) is not identical to computing the confidence-based ROC.

Binary ROC Data. The most common way to analyze sequential ROC data is to count any ID that occurs with a confidence rating greater than 0. If a filler ID with a confidence rating greater than 0 occurs first, the rest of the trial is effectively canceled, so neither the innocent suspect nor the guilty suspect will be identified by that witness. If a suspect ID occurs first, then it is counted as a hit if the suspect is guilty and counted as a false alarm if the suspect is innocent. The total number of target-present and target-absent lineups can be denoted *nTP* and *nTA*, respectively. In our study, *nTP* = 3258, and *nTA* = 3272. Across all lineups, the total number of hits and false alarms can be denoted *nH* and *nFA*. From these values, an overall (binary) hit and false alarm rate (HR and FAR, respectively) can be computed, where *HR* = *nH* / *nTP* and *FAR* = *nFA* / *nTA*. In our study, *nH* = 1687, so *HR* = 1687 / 3258 = .518. Because we did not have a designated innocent suspect in target-absent lineups, we estimated *nTA* (the number of false IDs of innocent suspects) by dividing the total number of first foil IDs in target-absent lineups (2229) by the lineup size of 6, such that *nTA* = 2229 / 6 = 371.5. Thus, *FAR* = 371.5 / 3272 = .114. Approximately the same FAR would be obtained if we instead randomly selected one foil on each target-absent lineup to serve as the designated innocent suspect. This HR, FAR pair (.518, .114) constitutes one point on the binary ROC plot, the one that corresponds to a neutral response bias. However, as noted earlier, one can use a more liberal or a more conservative criterion for counting IDs. We next reanalyzed the data using a liberal criterion by counting an ID if it was accompanied by a rating greater than -80. Thus, for example, if a face was rejected with a confidence rating of -70, it was now scored as a positive ID, but if a face was rejected with a rating of -90, it was still scored as a non-ID. Using this liberal decision rule allowed us to compute a second pair of overall hit and false alarm rates (i.e., a second binary ROC point). Finally, we reanalyzed the data again, this time under a conservative criterion – only counting an ID if it was accompanied by a rating of more than 80 (yielding a third binary ROC point). The binary ROC data are presented in Figure 7A. It is clear that they closely resemble the data generated earlier using a signal detection model (Figure 4B).

These are the first empirical data that correspond to the non-monotonic sequential ROC predicted by Rotello and Chen (2016). Thus, the results lend validity to their signal detection-based analysis. However, we did not actually manipulate the decision criterion across conditions (e.g., by using conservative, neutral or liberal instructions). If the overall decision criterion were successfully manipulated in that manner, there is no reason to expect that the results would differ in any appreciable way, but a definite answer would have to await further research. Keep in mind that the way we did choose to manipulate the overall decision criterion is an approach that is available to the police to use, should they wish to. For example, a particular jurisdiction could choose to adopt the sequential lineup procedure and to only count IDs made with relatively high confidence in an effort to reduce false IDs.

Confidence-based ROC data. For each of the overall criterion settings used for the binary analysis, we can also plot a confidence-based ROC, as we did earlier using simulated data based on a simple first-above-criterion signal detection model (Figure 4C). Figure 7B displays the empirical data computed the same way, and the predicted pattern was largely confirmed. Note that the neutral condition corresponds to how sequential ROC data have been plotted and analyzed in the past. However, although it seems natural, there is nothing inherently special about this particular criterion placement. If we wanted to know what empirical discriminability would be for identifications made with only higher levels of confidence, we could count only positive identifications that had been made with a confidence level of (for example) > 80. For a more liberal criterion, we could create a liberal curve by counting any decision made with a confidence level of > -80. Again, normally (i.e., with simultaneous lineups or showups), this approach would simply yield another point on the same ROC curve generated using a neutral response criterion. What is strange about the sequential lineup, however, is that the specific criterion placement changes whether or not one counts IDs made later in the lineup. As responding becomes more liberal, eyewitnesses increase the chances that they will make an incorrect identification on an earlier foil and not be permitted to identify a suspect that appears later in the lineup. This is why the three confidence-based ROCs shown in Figure 7B differ from each other (though the neutral and conservative ROCs are close).

The data shown in Figure 7B largely correspond to what we predicted using a simple signal detection model (Figure 4C), though the data from the neutral and conservative analyses are closer than expected. These results raise the possibility that, unlike with the simultaneous lineup, if two studies happened to differ only with respect to how liberal or conservative the participants were, their confidence-based ROC data would fall on different curves for that reason alone (i.e., even if underlying *d'* were the same in both studies).

Sequential Position Effects. Figure 8A shows the confidence-based sequential ROC data using the neutral decision criterion computed separately by position. Clearly, as predicted earlier using a simple signal detection model (Figure 6), the area under the ROC steadily decreases as position increases. To compare empirical discriminability across positions, *partial* area under the curve (pAUC) analyses must be conducted because none of the ROCs cover the full range of false ID rates. A common approach is to compare the pAUCs over the false ID rate range covered that is in common between the two conditions (i.e., over the range covered by the shorter of the two ROCs). Using the false ID rate range covered by the position 6 data, the difference in empirical discriminability between position 1 (pAUC = .013) and position 6 (pAUC = 0.010) was significant, *D* = 2.02, *p* = .044. The effect is barely significant only because so much of the position 1 data is excluded from this analysis. For example, using the larger false ID rate range covered by the position 3 data, the difference between position 1 (pAUC = .060) and position 3 (pAUC = .050) was more convincingly significant, *D* = 2.56, *p* = .010.

Table 2 shows the overall *HR* and *FAR* (i.e., rightmost ROC point) and the DR for each position. As predicted earlier (Table 1), the DR remains essentially constant. This result illustrates this importance of not relying on the DR as a measure of discriminability. Had this been the primary dependent measure in our experiment, it would appear as though the sequential lineup does not suffer from position effects despite the existence of rather large position effects in the ROC data. In fact, the prior use of the DR may explain why the field once came to the apparently mistaken conclusion that the sequential procedure is not compromised by position effects (Lindsay & Wells, 1985).

Figure 8B shows that the effect of position is magnified using a liberal decision criterion (counting any ID greater than -80). When a liberal criterion is used, most of the IDs will occur in position 1 (whether it is a foil or a suspect). Because almost all suspects appearing in position 1 will be identified, the hit and false alarm rates approach the upper right corner of the ROC (as they would in a showup). However, because foils appearing in position 1 will also almost always be identified, there are very few opportunities to identify suspects in positions 2 through 6; hence the dramatic drop in the ROCs for those positions.

Finally, as predicted in the signal detection-based simulations we presented earlier, position effects are minimized when a conservative criterion is used—that is, by counting IDs only if confidence exceeded 80 (Figure 8C). Note that the sequential procedure often does induce a conservative overall response bias. For example, in Experiment 1a of Mickes et al. (2012), the overall false alarm rate for the sequential procedure was .049. Thus, in many experiments, substantial position effects might not be expected. Here, however, the overall false alarm rate was .11 (i.e., responding was relatively liberal), which permitted large position effects to emerge. It is not clear why studies differ in the overall false alarm rate, but the fact that they do suggests that their empirically-measured discriminability results (in terms of partial area under the ROC) could differ in substantial ways for that reason alone.

Basic Model Fits. In an earlier investigation of sequential position effects, Horry, Palmer and Brewer (2012) found that when participants were unaware of how many photos would be shown (their "backloading" conditions), not only did the DR remain constant for positions 2 vs. 6 (see their Table 1), as we found here (Table 2), but so did underlying psychological discriminability (*d'*). However, instead of performing ROC analysis, they computed *d'* from a point estimate of the overall correct and false ID rates using a version of signal detection theory that assumes an "integration" decision rule (Duncan, 2006). The integration model holds that participants choose the MAX face in the lineup if the *sum* of the memory signals generated by the 6 faces in the line exceeds a decision criterion. Recently, Wixted et al. (2018) found that the integration model generally provides an extremely poor fit to empirical ROC data from simultaneous lineups (see also Colloff, Wade, Strange & Wixted, 2018), and they recommended that the field finally abandon that model. Moreover, as noted by Kaesler, Semmler, and Dunn (2017), another problem with the integration model, as it has been used in the past, is that it is not cognizant of the stopping rule and so may not provide a viable estimate of *d′* for that reason alone (i.e., even if the integration decision rule were viable, which it does not appear to be).

We next fit a signal detection model to the full set of data, collapsed across position. Importantly, we did not fit the Independent Observations model, according to which the face that generates strongest memory signal is identified so long as it exceeds the decision criterion (a model that makes sense for the simultaneous procedure). Instead, we fit the first-above-criterion model specified by Kaesler et al. (2017). This model is the appropriate signal detection model for a sequential lineup in which only the first positive ID counts. This model has up to 7 free parameters: μ_{Target}, σ_{Target} (both illustrated in Figure 1), and 5 confidence criteria (c_1 through c_5), corresponding to IDs made with confidence ratings > 0, > 20, > 40, > 60, and > 80. By convention, μ_{Foil} and σ_{Foil} were set to 0 and 1, respectively. The model was fit to the data using maximum likelihood estimation. We fit the model twice, first assuming a 6-parameter equal-variance model and then allowing for σ_{Target} to differ from σ_{Foil}. Goodness-of-fit was quantified by computing a χ^2 value for the observed and predicted IDs for each level of confidence (and for observed and predicted non-IDs) for both target-present and target-absent lineups. The results are shown in Table 3. Clearly, an unequal-variance model (χ^2 = 24.0) fit the data much better than an equal-variance model (χ^2 = 107.5), $\chi^2(1)$ = 83.5, p < .001, with the standard deviation of the target distribution estimated to be *less* than that of the lure distribution (i.e., $\sigma_{Target} < 1$). This is in contrast to what is commonly observed in studies of list memory, where σ_{Target} is usually greater than 1.0, with a typical value being 1.25 (e.g., Egan, 1958; Ratcliff, Shue, & Gronlund 1992; Wixted, 2007).

Why would the model yield an estimate of σ_{Target} less than 1? One possibility is that it reflects the fact that although every subject saw the same target (namely, a photo of the person seen in the mock crime video), the fillers were randomly drawn from a large pool of description-matched photos. Thus, the lure distribution, but not the target distribution, included item variance. It seems reasonable to suppose that selectively adding item variance to the fillers would result in a foil distribution with greater variance than would otherwise be the case. With enough item variance, the variance of the lure distribution would exceed the variance of the target distribution (as suggested by the best-fitting model). Yotsumoto, Kahana, McLaughlin, and Sekuler (2008) reported a similar result in a working memory task for visual textures. The task they used has some similarities to the task we used in that a small number of items were presented at study and test. Although their targets and foils both varied across trials, they proposed a summed-similarity account of recognition (implemented within a signal detection framework) that predicted that the variance of the memory-match signal generated by targets should be less than that generated by the foils. Thus, conceivably, the effect we observed here would be observed even if the target were varied across participants (i.e., even if every participant watched a different mock-crime video).

We next fit the model again, this time analyzing the data separately by position so that we could estimate whether μ_{Target} changes as a function of position. As shown earlier in Figure 8A, the area under the ROC decreases substantially with increasing position. Is the same true of μ_{Target}, which is a discriminability parameter (equal to *d'* in the equal-variance case)? We first fit a model in which μ_{Target} was fixed across positions and then compared it to a model in which μ_{Target} could differ between position 1 ($\mu_{Target1}$), effectively a showup, and the other six positions, for which μ_{Target} was equated ($\mu_{Target2\text{-}6}$). The results (Table 4) demonstrated that the fit was significantly improved by allowing μ_{Target} to differ in this way, $\chi^2(1) = 231.9 – 211.5 = 20.4$, $p < .001$. Note that when discriminability was allowed to change as a function of sequential position, $\mu_{Target2\text{-}6}$ (1.60) was greater than $\mu_{Target1}$ (1.37), which means that underlying discriminability increased for the later sequential positions even though empirical discriminability (the area under the ROC curve) decreased. This result is consistent with diagnostic feature-detection theory (Wixted & Mickes, 2014) described earlier. However, the fit was not further improved by allowing μ_{Target} to differ for positions 2 through 6. Diagnostic feature-detection theory would naturally predict an ever-increasing effect as participants come to better appreciate which facial features are non-diagnostic. These results indicate either that participants fully appreciated which features are non-diagnostic upon realizing that the face in position 2, like the face in position 1, also resembles the perpetrator or that some other explanation for the apparent increase in underlying discriminability applies. In any case, the data show no evidence of additional gains beyond position 2.

We next allowed σ_{Target} to also change between position 1 vs. positions 2-6 instead of holding it constant. Adding this parameter significantly improved the fit, $\chi^2(1) = 211.5 – 204.6 = 6.9$, $p = .009$, and both μ_{Target} and σ_{Target} were found to increase as a function of position. Their estimated values were 1.42 and 0.62, respectively, for position 1, and 1.60 and 0.78, respectively, for positions 2-6. Using these values to compute d_e (a d'-like discriminability measure that takes into account unequal variances) yields 1.58 for position 1 and 1.70 for positions 2 through 6 (Macmillan & Creelman, 2005). Using d_a (another d'-like discriminability measure that takes into account unequal variances) yields 1.71 for position 1 and 1.78 for positions 2 through 6.

For the analyses discussed thus far, the five confidence criteria were fixed across the six positions, but it seems possible that, in truth, they would differ across positions. We therefore next allowed the five confidence criteria to shift in lockstep for each of the six positions. This analysis added five additional parameters, one for each position 2 through 6. Thus, for example, if the position 2 shift parameter were estimated to be 0.05, it would mean that all of the confidence criteria shifted 0.05 standard deviations in the conservative direction for position 2 relative to position 1. Similarly, if the position 3 shift parameter were estimated to be -.05, it would mean that all of the confidence criteria shifted .05 standard deviations in the liberal direction for position 3 relative to position 1. We first performed this analysis with μ_{Target} and σ_{Target} fixed across positions. As shown earlier in Table 4, with μ_{Target} and σ_{Target} and the confidence criteria fixed across positions, the goodness-of-fit was $\chi^2(83) = 231.9$. When lockstep shifts of the confidence criteria was permitted, the fit of the model, $\chi^2(78) = 195.6$, was significantly improved, $\chi^2(5) = 231.9 – 195.6 = 36.3$, p < .001. The estimated criterion-shift parameters were 0.04, -0.09, -0.12, -0.02, and 0.03 for positions 2 through 6, respectively. Thus, although they shifted to a significant degree, the confidence criteria did not appear to shift in a very systematic way. When we next allowed μ_{Target} to vary for position 1 vs. positions 2-6, the fit was again significantly improved, $\chi^2(1) = 15.8$, $p < .001$, with its value being higher for position 1 (1.61) relative to the later positions (1.38). As before, allowing σ_{Target} to also vary as a function of position improved the fit still further, $\chi^2(1) = 7.4$, $p = .007$. Using the estimated values of μ_{Target} and σ_{Target} to compute d_e and d_a yielded virtually identical values as those reported above for the fixed-criteria analysis.[4]

Model-free estimates of underlying discriminability. In all of the previous analyses, the ROC curves for later positions were dragged down because participants who made an ID early

[4] We also allowed the confidence criteria to vary independently across positions, using 30 confidence parameters in all (5 for each position). The fit was significantly improved by the addition of these parameters, but no conclusions were affected. This version of the model is likely over-parameterized.

did not have the opportunity to make either a hit or a false alarm, but they still contributed to the denominator of the correct ID rate and the false ID rate. For applied purposes, all participants need to be included in the denominator because they are eyewitnesses who would be tested but would not have the opportunity to make a suspect ID. However, for theoretical purposes, it is worth separately considering the performance of only the subset of participants at each position who have not yet eliminated themselves because of a prior filler ID. For example, imagine that 600 participants were presented with a target-present lineup containing the guilty suspect in position 3. If 200 made a previous filler ID, 200 others did not make a previous filler ID and made a correct suspect ID, and the remaining 200 did not make any ID in the first three positions, the position 3 hit rate would be 200 / 600 = .33. However, because of the stopping rule, only 400 participants had an opportunity to identify the suspect in position 3. For that subset of participants considered separately, the hit rate would be 200 / 400 = .50. Plotting the ROC data in this manner effectively removes the fundamental constraint imposed by the stopping rule and would allow any increase in discriminability as a function of position to be observed without having to fit a model.

Using this approach, the area under the empirical ROC should now reflect underlying discriminability. Figure 9 shows the ROC curve for position 1 plotted against the ROC curve for positions 2-6 when considering only the subset of participants who have not yet made an identification prior to getting to each position. The ROC is visually lower for position 1 than it is for the later positions (consistent with the model fits).

We also performed this model-free area-under-the-curve analysis separately by the sequential position of the suspect. Specifically, we computed pAUC for each of the six ROCs computed in the manner described above, using the false alarm rate range covered by the position-6 ROC (a range over which all six ROCs yielded hit- and false-alarm rate data). Figure 10 shows the results. The figure also shows the mean pAUC (depicted by the solid horizontal line) and 95% confidence interval (depicted by dashed horizontal lines) computed using the pAUC values for positions 2 through 6. The results of this analysis correspond to the analyses reported thus far. That is, discriminability was higher (and essentially constant) for positions 2 through 6 relative to position 1, the discriminability of which falls outside the range of the subsequent values. At a minimum, without relying on any model fitting, this analysis shows that psychological discriminability does not decrease as a function of sequential position – if anything, it increases, despite the fact that empirical discriminability (computed using the stopping rule) decreased dramatically over the same range (Figure 8A).

Reaction time data. The apparent increase in discriminability as a function of sequential position was predicted by diagnostic feature-detection theory, but it is obviously not the only possible explanation. Another possible explanation for the higher psychological performance for later positions is that participants change their speed-accuracy boundaries over the course of testing. According to the diffusion model, spreading out the decision boundaries would increase accuracy and also result in longer reaction times (Ratcliff, 1978). Table 5 shows median response times for the decisions made in each position. The data show that response times were longer for the first position than for later positions such that, if anything, changing decision boundaries worked against detecting an increase in discriminability as sequential position increased. In target present lineups, response times were significantly longer for position 1 than for position 6, $t(3187) = 9.85$, $p < .001$. In target absent lineups, response times were also significantly longer for position 1 than for position 6, $t(3171) = 20.80$, $p < .001$. Note that these are within-subject *t*-tests because each participant made a response in position 1 (whether “yes” or “no”) and in position 6.

We also performed an analysis of response times for the first “yes” decision to the suspect in target-present lineups and for the first “yes” decision to a filler in target-absent lineups. These are the “yes” responses used for the model-based analyses of underlying discriminability reported above (which show that, if anything, discriminability increased with later suspect positions). Figure 11 shows the relevant data. In both Figure 11A (target-present) and 11B (target-absent), the response time values for positions 2 through 6 were used to compute the average of the median response times (depicted by the solid horizontal line) and 95% confidence interval (depicted by dashed horizontal lines). In both cases, the median response time for position 1 fell above the range of the subsequent response times. Thus, we infer that the increased in underlying discriminability for positions 2-6 relative to position 1 did not occur because participants allowed more information to accumulate before responding in the later positions.

Sequential dependencies. The signal detection model we fit to the data does not take into consideration the possibility that how participants responded on earlier positions influenced how they responded on later positions. For example, once participants make an ID, they may be hesitant to make a subsequent ID because they think they have already identified the guilty person, and there can be only one guilty person. We explored this possibility by examining the average confidence rating to fillers before and after a target was identified in target-present lineups and to fillers before and after another filler was identified in target-absent lineups. In other words, we examined how responding changed immediately after making a first positive identification.

This analysis revealed the existence of sequential dependencies. That is, in both target-present and target-absent lineups, participants were more likely to declare that a subsequent filler was not the perpetrator (i.e., they became more conservative) after making an identification than before making an identification. Table 6 shows the average confidence ratings given to the filler immediately before and immediately after responding “ID” or “No ID.” The difference between the before and after ratings differed significantly depending on whether a suspect was identified or not identified, $t(1288) = 5.09$, $p < .001$. The difference between the before and after ratings also differed significantly depending on whether a filler was identified or not identified, $t(7841) = 4.69$, $p < .001$. Thus, not surprisingly, participants effectively became more conservative after making a first identification.

The signal detection model we used predicts that the first-identification-only rule itself is what creates the empirical position effects we observed (i.e., decreasing area under the ROC with increasing sequential position). According to this model, if the highest-confidence ID is counted instead of only counting the first ID, these artificially-induced position effects should no longer be apparent (i.e., the ROCs from each position would fall atop one another). However, as just noted, due to sequential dependencies in responding, participants were less likely to make a subsequent identification after making an initial identification. This suggests that position effects similar to those created by the first-identification-only rule may still be present even when the stopping rule is eliminated and each participant’s highest-confidence ID is counted as the ID. Indeed, Figure 12 shows the same general pattern of position effects when the highest-confidence ID is counted rather than counting the first-ID. Position effects are clearly attenuated compared to when the stopping rule is in effect, but a self-imposed stopping rule (i.e., reluctance to make another ID if one was already made) still creates clear position effects, with empirical discriminability being lower for the later sequential positions. Partial area under the curve analysis over the false ID rate range covered by the position 6 data revealed that the difference between position 1 (pAUC = 0.033) and position 6 (pAUC = 0.026) was significant, $D = 2.53$, $p = .012$. We found essentially the same pattern if each position is treated as a showup (counting suspect IDs regardless of whether a previous ID was made).

Discussion

The structural constraints of the sequential procedure that are imposed by the stopping rule create a situation wherein *empirical* discriminability (i.e., the area under the ROC curve) differs from *underlying* discriminability. Having seen a previous face improves eyewitnesses' ability to discriminate innocent from guilty suspects. This can be best appreciated when the data are fit using a model that is cognizant of the underlying structure of the test. The observation that underlying discriminability is higher for later positions, however, does not mean that police who use a sequential lineup with a first-identification-only rule should place the suspect in a later lineup position. In fact, the recommendation to policymakers that follows from this experiment would be exactly the opposite because empirical discriminability can drop off dramatically as the suspect appears in a later lineup position.

No matter how the data are analyzed, the results indicate that underlying discriminability was higher for suspect positions 2 through 6 relative to suspect position 1, a finding that was predicted by diagnostic feature-detection theory. However, the results do not offer definitive support of this theory. First, even though the increase in discriminability was apparently real, the effect was very small and likely would not even be detected unless a large number of participants were tested (as they were in this experiment). Second, the most straightforward prediction based on this theory is that discriminability should increase in continuous fashion with each sequential suspect position as participants accumulate knowledge about which features are non-diagnostic.

Yet, a step function was observed instead. As such, it seems possible that the increase in discriminability for the later positions reflects an orientation effect. That is, conceivably, participants were both slow and diagnostically inefficient when making a decision about the first face in the lineup, but they were better oriented to the task after that (and so responded more efficiently and effectively over the subsequent positions). Whatever the explanation, the key point is that underlying discriminability was higher for the later suspect positions, in sharp contrast to empirical discriminability, which exhibited the opposite effect.

Although underlying discriminability apparently increases for later lineup positions, the fact that empirical discriminability decreases by position suggests that a sequential lineup might be expected to perform worse than a showup. Fundamentally, a showup is a sequential lineup with the suspect in the first position. The only real difference between the two procedures is that when eyewitnesses see the face in the first position of a sequential lineup, they know they will view additional faces afterwards, whereas with a showup, eyewitnesses know that they will view only a single test face. Showups are often called suggestive because they “suggest” that the face being tested is likely to be the guilty suspect (Dysart & Lindsay, 2012). Therefore, showups would be expected to result in more liberal responding than sequential lineups, but using the sequential procedure (rather than a showup) may not enhance empirical discriminability. In fact, as noted above, the countervailing force of increased underlying discriminability for later suspect positions is quite small and therefore seems unlikely to appreciably counteract the more pronounced negative force exerted by the stopping rule on empirical discriminability as suspect position increases. In our second experiment, we directly compared a showup to the sequential lineup.

Experiment 2

Method

Participants

A total of 4398 subjects completed the task on MTurk, but some answered the attention-check question incorrectly and were excluded, leaving a total of 3919 participants. Of those, 1966 were randomly assigned to the sequential lineup condition, (999 were presented with a target-present lineup and 967 with a target-absent lineup) and 1953 were randomly assigned to the showup condition (980 were presented with a target-present showup and 973 with a target-absent showup). Participants were each paid $0.25 for completing the task.

Materials

The materials were identical to those used in Experiment 1.

Procedure

For the participants tested with a sequential lineup, the procedure was exactly the same as it was in Experiment 1. For the participants tested with a showup, the study phase and distractor task were the same, except that they were informed they would see only a single face at test.

Results

Figure 13 displays the empirical ROC curves for the showup vs. the sequential lineup (collapsed across position). Note that the showup curve represents positive IDs only (i.e., the curve could have been extended to the upper right corner by including showup rejections, as is typically done with old/new ROC data). Obviously, the sequential lineup did not yield a higher area under the curve than the showup. If anything, the trend is slightly in the opposite direction, though the partial area under the curve over the range covered by the sequential procedure (namely, FAR = 0 to FAR = .114) did not differ significantly from that of the showup. It is apparent in Figure 13 that the overall hit and false alarm rates for the showup considerably exceed the corresponding values for the sequential lineup (rightmost ROC point for each procedure). This is consistent with the idea that showups are “suggestive.” For the showup, the overall hit and false alarm rates were .863 and .255, respectively (DR_{Showup} = 3.39). The corresponding values for the sequential procedure were .518 and .114, respectively ($DR_{Sequential}$ = 4.56). Thus, using the DR to gauge the diagnostic accuracy of an identification procedure, as researchers did for many years, one would judge the sequential lineup to be superior to the showup. However, according to these data, the use of a showup in conjunction with a conservative decision criterion could achieve a lower false alarm rate and the same (if not higher) hit rate compared to the sequential procedure.

Then again, the sequential ROC curve is affected by the overall decision criterion, which was relatively liberal in this experiment. The more liberal the overall decision criterion, the lower the confidence-based ROC for the sequential procedure will be. By contrast, the showup is represented by a single ROC curve, because it follows the standard rules wherein becoming more liberal or conservative in responding simply creates another point on the *same* ROC curve. How would the showup ROC compare to the lineup ROC if a more conservative decision criterion had been used?

Before addressing that question, Figure 14 presents the empirical ROC curves for three different decision criteria, with the binary ROC shown in Figure 14A for three different overall decision criteria (> -80, > 0, and > 80), and the corresponding confidence-based ROCs shown in Figure 14B. These data largely replicate the findings of Experiment 1.

Figure 15 reproduces the data in Figure 14B along with the showup data shown earlier in Figure 13. Interestingly, and as would be expected, if responding is conservative enough, the points on the sequential ROC overlap with the points for the showup. Thus, at least according to this analysis, sequential lineups can certainly achieve more conservative responding and a higher DR than showups computed from overall hit and false alarm rates, but they may not necessarily achieve higher empirical discriminability when a fair lineup is used.

Model fits. In terms of underlying discriminability, the showup and overall sequential lineup yielded similar results. As shown in Table 7, μ_{Target} was nearly identical for the two procedures, and σ_{Target} less than 1 in both cases. When μ_{Target} and σ_{Target} were constrained to be equal for the two procedures, the overall chi-square goodness-of-fit statistic did not increase significantly, $\chi^2(2) = 4.20$, $p = .122$. In addition, when we fit the sequential data separately by position, there were no significant differences between position 1 and the later positions in terms of μ_{Target}. Note that Experiment 2 was not designed to detect significant position effects, whereas Experiment 1 was, using more than three times as many participants. Experiment 2 was designed to compare the sequential procedure to the showup procedure and may not have had the statistical power to detect significant position effects on *d'* that might exist.

General Discussion

The sequential lineup is widely used by the police in the U.S., but it is not well understood theoretically. The theory most often associated with the sequential procedure holds that it promotes “absolute” decisions by focusing attention on a face presented in isolation instead of encouraging witnesses to compare that face to the other faces in the lineup (as a simultaneous procedure does). As originally conceived (Wells, 1984), this was a theory of response bias. That is, theoretically, the simultaneous presentation of faces biases eyewitnesses to choose the most familiar face in the lineup even if that face does not match the witness’s memory of the perpetrator very well. By contrast, the presentation of faces in isolation theoretically reduces that bias to choose, resulting in lower hit and false alarm rates (and increasing the DR). However, because response bias can be easily manipulated without switching lineup procedures, a more interesting question concerns the effect of sequential lineups on discriminability compared to other identification procedures.

Empirical vs. Underlying (Psychological) Discriminability. Previous studies comparing simultaneous lineups, sequential lineups, and showups have generally found that the partial area under the ROC curve (i.e., empirical discriminability) is greatest for simultaneous lineups (Mickes & Gronlund, 2017). Diagnostic feature-detection theory (Wixted & Mickes, 2014) holds that these findings reflect the advantage conferred by simultaneous lineups with respect to the detection of non-diagnostic facial features. In other words, it holds that empirical discriminability is higher for simultaneous lineups compared to sequential lineups and showups because underlying psychological discriminability is higher for simultaneous lineups. However, the results reported here show that structural constraints associated with the sequential presentation of faces can reduce the area under the ROC curve for that procedure even if psychological discriminability (underlying *d'*) is unaffected, as predicted by Rotello and Chen (2016).

Our findings provide a clear demonstration that psychological and empirical discriminability need not agree with each other (Wixted & Mickes, 2018). In fact, they can go in opposite directions, as they did here. Earlier, we showed that the simplest signal detection model of sequential lineup performance predicts that empirical discriminability (pAUC) should decrease as the position of the suspect in the sequential lineup increases (Figure 6A) even if underlying discriminability (*d'*) remains constant across suspect positions. We confirmed that prediction in Experiment 1 (Figure 8A). At the same time, diagnostic-feature-detection theory predicts that underlying discriminability should *increase* as the position of the suspect in the sequential lineup increases. Psychological discriminability is predicted to increase because seeing prior faces should teach eyewitnesses that the faces in the lineup share features. Because such features are non-diagnostic (precisely because they are shared by everyone in the lineup), discounting them enhances the ability to discriminate between innocent suspects/fillers and guilty suspects. One way to detect increased discriminability by position, if it exists, is to fit a model to the data that is cognizant of the structural constraint imposed by the stopping rule in the sequential procedure (Kaesler et al., 2017). Indeed, when such a model is fit to our data, the results show that psychological discriminability increased as a function of suspect position (Table 4) despite the *decreasing* empirical discriminability as a function of suspect position.

The increase in underlying psychological discriminability with increasing suspect position was fairly small and likely would not have been detected had we not tested a large number of participants in Experiment 1. Indeed, the effect was not detected in Experiment 2, which tested only about one-third the number of participants tested using the sequential procedure in Experiment 1. Although small, the effect seems real because it is apparent no matter how the data are analyzed, including using a model-free method where pAUC ought to reflect underlying psychological discriminability because the constraint imposed by the stopping rule was removed (e.g., Figure 10). We investigated the effect of suspect position on underlying discriminability because it was predicted by diagnostic feature-detection theory (Wixted & Mickes, 2014). Although the data support that theory, the small size of the effect may indicate that the discounting of non-diagnostic features is much less pronounced when faces are presented sequentially compared to when they are presented simultaneously. Alternatively, given the step-function nature of the increase in psychological discriminability from position 1 to positions 2-6, the results may not reflect the discounting of non-diagnostic features at all and may simply indicate that participants became better oriented to the task after making a decision to the first face in the lineup. Regardless of which explanation applies, the results demonstrate that empirical discriminability and underlying psychological discriminability can be affected in opposite ways.

"Filler siphoning" and the effect of response bias on the sequential ROC. Our main findings were based on analyses that assumed a neutral response bias, but we also investigated the effect of liberal and conservative response biases on the empirical ROC data generated by the sequential procedure. By artificially manipulating the overall decision criterion for counting IDs (as illustrated in Figure 5), we found that changing response bias can magnify or minimize effects of sequential position on empirical discriminability. When we used a liberal criterion, position effects were magnified, but when we used a conservative criterion, they were all but eliminated (Figure 8). Although we did not manipulate overall response bias across conditions using instructions, we predict that the same effects would be observed using that approach (to the extent that the instructions successfully manipulated response bias).

The fundamental constraint that reduces empirical discriminability for later suspect positions in the sequential lineup is that fillers that appear before the suspect in the lineup and that happen to generate an above-criterion memory signal effectively terminate the procedure, thereby preventing guilty suspect IDs that might have otherwise occurred. Expressed in language that is sometimes used in the eyewitness identification literature, the constraint imposed by the use of sequential lineups in conjunction with a stopping rule is caused by "filler siphoning" (Wells, Smalarz, & Smith, 2015; Smith, Wells, Lindsay, & Penrod, 2016). Filler siphoning refers to the fact that the presence of fillers in simultaneous or sequential lineups reduces the number of suspect IDs that would occur in their absence.

Filler siphoning is usually considered to be a beneficial phenomenon because, compared to a showup, fillers in a simultaneous lineup draw IDs away from innocent suspects to a greater extent than guilty suspects (e.g., Smith, Wells, Smalarz & Lampinen, 2018). This is another way of saying that filler siphoning has the effect of increasing the DR for lineups compared to showups, just as the use of a more conservative response criterion would (Colloff et al., 2018). Thus, as currently construed, the apparently beneficial effect of filler siphoning (namely, an increase in the DR) is exactly the same effect that was once thought to indicate a sequential superiority effect. However, just as a higher DR resulting from more conservative responding does not reflect superior empirical discriminability (e.g., in a comparison of relatively liberal simultaneous lineups vs. relatively conservative sequential lineups), a higher DR resulting from filler siphoning also does not reflect superior empirical discriminability (e.g., in a comparison of showups, which have no fillers, vs. simultaneous or sequential lineups, which do). As noted by Colloff et al. (2018), filler siphoning could increase the DR whether empirical discriminability, as measured by ROC analysis, increased, decreased or remained unchanged (just as is true of more conservative responding).

Not only is the effect of filler siphoning not necessarily beneficial in terms of empirical discriminability, it is clearly detrimental in the case of sequential lineups. If a filler happens to generate a relatively weak memory signal that barely surpasses a liberal decision criterion, the filler will be identified (presumably with low confidence), thereby canceling the opportunity to identify the guilty suspect in a later position, who might have generated a much stronger memory signal. The effect of filler siphoning, which exerts downward pressure on the empirical ROC for later sequential positions, can more than cancel any positive effect of increasing underlying psychological discriminability that might occur as a function of sequential position. That clearly occurred in our Experiment 1, where the empirical ROCs showed no sign of increasing discriminability as a function of suspect position (instead, it monotonically decreased) despite the fact that underlying discriminability increased. The detrimental effect of filler siphoning on empirical discriminability may explain why the sequential procedure did not outperform and instead slightly but non-significantly *underperformed* the showup procedure in Experiment 2 (Figure 13).

Interestingly, and somewhat surprisingly, for fair lineups of the kind we used here, there is no evidence anywhere in the scientific literature that sequential lineups yield higher empirical discriminability than showups. Sequential lineups clearly *do* elicit much more conservative responding than showups, thereby increasing the DR. In the past, this result has been mistakenly interpreted to mean that sequential lineups are superior to showups. However, as noted above, this result may instead simply reflect the fact that sequential lineups induce more conservative responding than showups, as was true in Experiment 2 here (see Figure 13). Indeed, the DR was substantially higher for the sequential procedure ($DR_{Sequential} = 4.56$ vs. $DR_{Showup} = 3.39$) even though empirical discriminability was slightly lower.

Unfair lineups reduce filler siphoning. The sequential procedure's filler-siphoning constraint occurs only to the extent that fillers generate a memory signal that exceeds the decision criterion. Having a conservative criterion decreases the chances that a filler will exceed the criterion, minimizing position effects on empirical discriminability (Figure 8C), but another factor that can reduce filler siphoning is the use of fillers that do not resemble the perpetrator. In the extreme, for example, if only the suspect (innocent or guilty) closely matched the description of the perpetrator, then filler IDs would be rare. Under such conditions, participants would more often have the opportunity to identify the guilty suspect, even if the suspect was positioned late in the lineup. Imagine, for example, that the mock-crime video depicted a perpetrator who was a clean-shaven White male in his early 20s with short dark brown hair. If the fillers generally matched that description except that they were all in their 50s, then they would provide some useful information about non-diagnostic features (e.g., they would reveal that short dark brown hair is not diagnostic of guilt), but they would be unlikely to be identified because they are too old. Thus, filler siphoning would be effectively eliminated. Under such conditions, any increase in underlying psychological discriminability might show up as an increase in empirical discriminability for the later suspect positions.

These considerations may help to explain why two previous studies have reported significantly increased empirical discriminability for later sequential lineup positions (Meisters, Diedenhofen & Musch, 2018; Gronlund et al., 2012). Both of these studies used unfair lineups in which fillers were less likely to be identified than the designated innocent suspect in the target-absent lineup. Thus, the fillers were less likely to generate a memory signal strong enough to exceed the decision criterion (as illustrated in Figure 16). Using 6-person sequential lineups, Gronlund et al. (2012) found that pAUC for position 5 (0.141) significantly exceeded pAUC for position 2 (0.092). This increase in empirical discriminability presumably reflects a corresponding increase in underlying psychological discriminability that was not counteracted by filler siphoning. The same result was reported in a recent study by Meisters, Diedenhofen and Musch (2018), who also tested unfair lineups in which filler IDs were rare. In their 4-person sequential lineups, pAUC for position 4 (0.10) was significantly higher than it was for position 1 (0.05). By contrast, in fair lineups of the kind we tested here (where filler siphoning would be expected to occur and in fact did frequently occur), any increase in underlying psychological discriminability as a function of suspect position is apparently swamped by the negative effect of filler siphoning on empirical discriminability.

Because underlying psychological discriminability (measured by *d′*) and empirical discriminability (measured by pAUC) need not necessarily agree with each other, it is important to know which measure answers the question of interest (Wixted & Mickes, 2018). Both types of discriminability are important, but it is essential to appreciate *when* each is important. When the question pertains to a prediction made by a theory, such as diagnostic feature-detection theory, a model is needed to measure underlying psychological discriminability (*d′*). But when the question concerns application in the real world, only empirical discriminability, measured by pAUC, is relevant (Wixted & Mickes, 2018).

Applied Considerations. The idea that underlying psychological discriminability is of no interest to policymakers is not intuitive, but the data we report here illustrate why it is so. For example, if a suspect is placed in position 6 of the sequential procedure and a stopping rule is used, it would be no consolation to know that psychological discriminability may be higher for that position compared to position 1. The fact that empirical discriminability is low for the last position means that, for applied purposes, routinely placing suspects in position 6 would be a bad idea. For applied purposes, the goal of any eyewitness identification procedure is to maximize *empirical* discriminability. Higher empirical discriminability both reduces the likelihood that innocent suspects will be misidentified (and possibly wrongfully convicted) and increases the likelihood that guilty suspects will be prevented from committing future crimes.

Remarkably, approximately 30% of more than 15,000 U.S. police departments have adopted the sequential lineup procedure (Police Executive Research Forum, 2013) even though it is not yet well understood at a theoretical level. Only recently have researchers begun to understand the fundamental constraints on empirical discriminability created by the sequential procedure. Our findings provide an empirical demonstration of these constraints. Rotello and Chen (2016) were the first to theoretically identify the constraint, and it was clearly evident in our empirical data even when a neutral decision criterion was used (Figure 8A). Using the same basic signal detection model that they did, we predicted (Figure 6) and documented (Figure 8) further constraints on empirical discriminability as a function of the suspect's position in the sequential lineup. These findings seem significant because, in the original study that introduced the sequential procedure, Lindsay and Wells (1985) stated that "[f]or sequential lineup presentation to be a viable alternative [to simultaneous presentation], it is important that the results of the procedure not be unduly influenced by order effects (i.e., the position of the suspect)" (p. 561). The present research demonstrates how large an influence the position of the suspect can have on discriminability and how that effect can be masked by relying on a dependent measure like the DR (e.g., see the DR values in Table 2, which remain essentially constant as a function of suspect position).

Most of the problems with the sequential procedure arise because of the standard stopping rule used in laboratory research. It is not clear how often police use this stopping rule in actual practice (e.g., Steblay et al., 2011, noted that they are unaware of any jurisdiction that does), but given its deleterious effect on empirical discriminability, the police would probably be wise not to follow the first-identification-only stopping rule. However, even if they do not use the stopping rule, the data shown in Figure 12 (where the highest-confidence ID was used rather than the first ID) suggest that sequential lineups can still lead to deleterious position effects on empirical discriminability. If fair sequential lineups do not yield higher discriminability than showups, and if they yield lower empirical (not to mention psychological) discriminability than simultaneous lineups, the argument in favor of police switching to the sequential procedure is hard to fathom. The larger implication may be that before advocating major changes to existing policy, scientists should have a deep theoretical understanding of any proposed reform, one that is as grounded in basic research as it is in applied research.

References

Benjamin, A. S., Tullis, J. G., & Lee, J. H. (2013). Criterion noise in ratings-based recognition: Evidence from the effects of response scale length on recognition accuracy. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*, 1601-1608.

Carlson, C. A., Gronlund, S. D., & Clark, S. E. (2008). Lineup composition, suspect position, and the sequential lineup advantage. *Journal of Experimental Psychology: Applied*, 14, 118 –128.

Colloff, M. F., Wade, Kimberley A., Strange, D. & Wixted, J. T. (2018) Filler siphoning theory does not predict the effect of lineup fairness on the ability to discriminate innocent from guilty suspects: reply to Smith, Wells, Smalarz, and Lampinen. *Psychological Science*. First Published August 3, 2018, https://doi.org/10.1177/0956797618786459

Criss, A. H., Malmberg, K. J., & Shiffrin, R. M. (2011). Output interference in recognition memory. *Journal of Memory and Language*, 64, 316–326.

Clare, J., & Lewandowsky, S. (2004). Verbalizing facial memory: Criterion effects in verbal overshadowing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30, 739 –755. http://dx.doi.org/ 10.1037/0278-7393.30.4.739

Dube, C., & Rotello, C. M. (2012). Binary ROCs in perception and recognition memory are curved. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*, 130–151.

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

Dysart, J. E., & Lindsay, R. C. L. (2012). Show-up identifications: Suggestive technique or reliable method? In *The Handbook of Eyewitness Psychology* (Vol. 2, pp. 137-153). New

York, NY: Routledge.

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

Goodsell, C. A., Gronlund, S. D., & Carlson, C. A. (2010). Exploring the sequential lineup advantage using WITNESS. *Law and Human Behavior,* 34, 445– 459. doi:10.1007/s10979-009-9215-7

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

Gronlund, S. D., Wixted, J. T., & Mickes, L.(2014). Evaluating eyewitness identification procedures using ROC analysis. *Current Directions in Psychological Science, 23*, 3–10. http://dx.doi.org/10.1177/0963721413498891

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

Kaesler, M. P., Semmler, C. & Dunn, J. (2017). Using Measurement Models to Understand Eyewitness Identification. 39th Annual Meeting of the Cognitive Science Society. G. Gunzelman, A. Howes, T. Tenbrik & E. Davelaar (Eds.) London, UK.

Koen, J. D., & Yonelinas, A. P. (2011). From humans to rats and back again: Bridging the divide between human and animal studies of recognition memory with receiver operating characteristics. *Learning & Memory, 18*, 519–522.

Lindsay, R.C.L., & Wells, G.L. (1985). Improving eyewitness identification from lineups:

Simultaneous versus sequential lineup presentations. *Journal of Applied Psychology*, *70*, 556–564.

Mickes, L. & Gronlund, S. D. (2017). *Eyewitness Identification*. J. H. Byrne (Ed.) Learning and Memory: A Comprehensive Reference 2E, Volume Cognitive Psychology of Memory. Elsevier.

Mickes, L., Seale-Carlisle, T.M., Wetmore, S.A., Gronlund, S.D., Clark, S.E., Carlson, C.A., Goodsell, C.A., Weatherford, D. & Wixted, J.T. (2017). ROCs in eyewitness identification: Instructions versus confidence ratings. *Applied Cognitive Psychology*, *31*, 467-477. doi: 10.1002/acp.3344

Murdock, B. B., & Anderson, R. E. (1975). Encoding, storage, and retrieval of item information. In R. L. Solso (Ed.). Theories in cognitive psychology: The Loyola Symposium. Erlbaum.

Osth, A. F., Jansson, A., Dennis, S., & Heathcote, A. (2018). Modeling the dynamics of recognition memory testing with an integrated model of retrieval and decision making. *Cognitive Psychology*, *104*, 106-142.

Peixotto, H. E. (1947). Proactive inhibition in the recognition of nonsense syllables. *Journal of Experimental Psychology*, *37*(1), 81–91.

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

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological review, 85*(2), 59.

Ratcliff, R., Sheu, C. F., & Gronlund, S. D. (1992). Testing global memory models using ROC curves. *Psychological Review, 99*, 518–535.

Rotello, C. M., & Chen, T. (2016). ROC analyses of eyewitness identification decisions: An analysis of the recent debate. *Cognitive Research: Principles and Implications*.

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

Smith, A. M., Wells, G. L., Lindsay, R. C. L., & Penrod, S. D. (2017). Fair Lineups Are Better Than Biased Lineups and Showups, but Not Because They Increase Underlying Discriminability. *Law and Human Behavior*, *41*,127-145.

Smith, A. M., Wells, G. L., Smalarz, L., & Lampinen, J. M. (2018). Increasing the similarity of lineup fillers to the suspect improves the applied value of lineups without improving memory performance: Commentary on Colloff, Wade, and Strange (2016). *Psychological Science*. First Published August 3, 2018, https://doi.org/10.1177/0956797617698528

Swets, J. A., Tanner, W. P., & Birdsall, T. G. (1961). Decision processes in perception. *Psychological Review, 68,* 301-340.

Wells, G. L., Smalarz, L., & Smith, A. M. (2015). ROC analysis of lineups does not measure underlying discriminability and has limited value. *Journal of Applied Research in Memory and Cognition*, *4*, 313–317. doi:10.1016/j.jarmac. 2015.08.008

Wilson, B. M., Seale-Carlisle, T. M., & Mickes, L. (2018). The Effects of Verbal Descriptions on Performance in Lineups and Showups. *Journal of Experimental Psychology: General, 147*, 113-124.

Wixted, J. T. (2007). Dual-process theory and signal-detection theory of recognition memory. *Psychological Review, 114*, 152-176.

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

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

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

Yotsumoto, Y., Kahana, M. J., McLaughlin, C., and Sekuler, R. (2008). Recognition and position information in working memory for visual textures. *Memory & Cognition*, *36*, 282–294.

Table 1

Overall hit and false alarm rates (i.e., the rightmost ROC points) and their corresponding diagnosticity ratios for the simulated ROC curves shown in Figure 6A.

Position	HR	FAR	DR
1	0.79	0.18	4.44
2	0.65	0.15	4.33
3	0.54	0.12	4.69
4	0.44	0.10	4.62
5	0.36	0.08	4.50
6	0.30	0.06	4.63

Table 2

Overall hit and false alarm rates (i.e., the rightmost ROC points) and their corresponding diagnosticity ratios for the empirical ROC curves shown in Figure 8A.

Pos	HR	FAR	DR
1	0.82	0.18	4.52
2	0.64	0.13	5.08
3	0.54	0.13	4.08
4	0.47	0.11	4.09
5	0.34	0.07	4.61
6	0.29	0.05	5.67

Table 3

Optimal parameter estimates for the fits of the equal-variance and unequal-variance signal detection models to the data (collapsed across positions) from Experiment 1.

Parameter	**Equal Variance**	**Unequal Variance**
μ_{Target}	1.63	1.56
σ_{Target}	0.71	0.71
c1	0.91	0.92
c2	0.94	0.95
c3	1.03	1.04
c4	1.23	1.22
c5	1.75	1.71
χ^2	107.5	24.0
df	9	8
p	0.000	0.002

Table 4

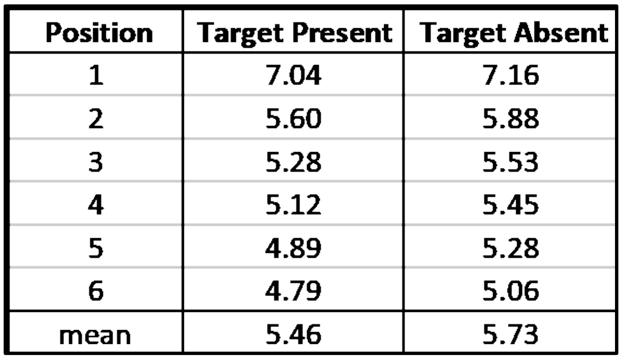
Optimal parameter estimates for the fits of the constant-discriminability and changing-discriminability signal detection models to the data (not aggregated across positions) from Experiment 1.

Parameter	**Constant Discriminability**	**Changing Discriminability**
$\mu_{Target1}$	1.54	1.37
$\mu_{Target2-6}$	1.54	1.60
σ_{Target}	0.73	0.73
c1	0.92	0.92
c2	0.95	0.95
c3	1.04	1.04
c4	1.22	1.22
c5	1.70	1.70
χ^2	231.9	211.5
df	83	82
p	0.000	0.000

Note. The degrees of freedom are large in this case because the model is fit to multiple data cells for each position.

Table 5

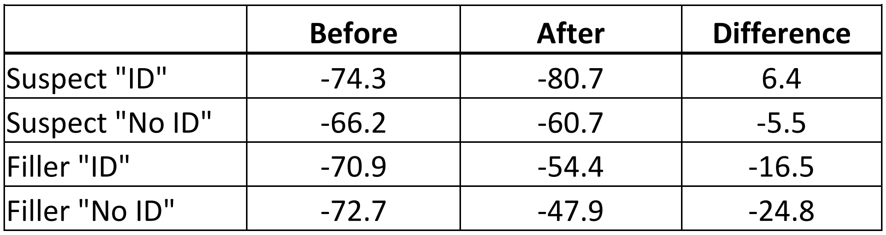
Average response times for each position in target present and target absent lineups.



Position	Target Present	Target Absent
1	7.04	7.16
2	5.60	5.88
3	5.28	5.53
4	5.12	5.45
5	4.89	5.28
6	4.79	5.06
mean	5.46	5.73

Table 6

Average confidence ratings given to the filler immediately before and immediately after responding “ID” or “No ID.”



	Before	After	Difference
Suspect "ID"	-74.3	-80.7	6.4
Suspect "No ID"	-66.2	-60.7	-5.5
Filler "ID"	-70.9	-54.4	-16.5
Filler "No ID"	-72.7	-47.9	-24.8

Table 7

Optimal parameter estimates for the fits of the unequal-variance signal detection models to the showup data and the sequential (SEQ) lineup data (collapsed across positions) from Experiment 2.

Parameter	**Showup**	**SEQ Lineup**
μ_{Target}	1.58	1.60
σ_{Target}	0.90	0.73
c1	0.62	1.03
c2	0.69	1.02
c3	0.86	1.11
c4	1.16	1.26
c5	1.85	1.69
χ^2	7.0	11.5
df	3	8
p	0.071	0.175

Figure 1. Simple equal-variance signal detection model. The distribution of foils has a lower memory match signal on average than the distribution of targets because the foils have not been previously seen, whereas the targets have been previously seen. The vertical line represents one of many possible places where the decision criterion is placed. An item generating a memory match signal greater than the criterion will be identified as "old;" an item generating a memory match signal lower than the criterion will be identified as "new."

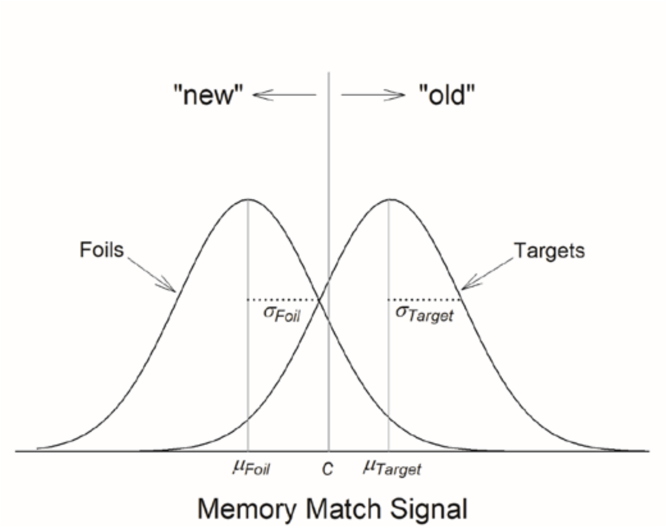


Figure 2. Confidence-based ROC data (A) and ROC data resulting from different biasing instructions in Mickes et al. (2017).

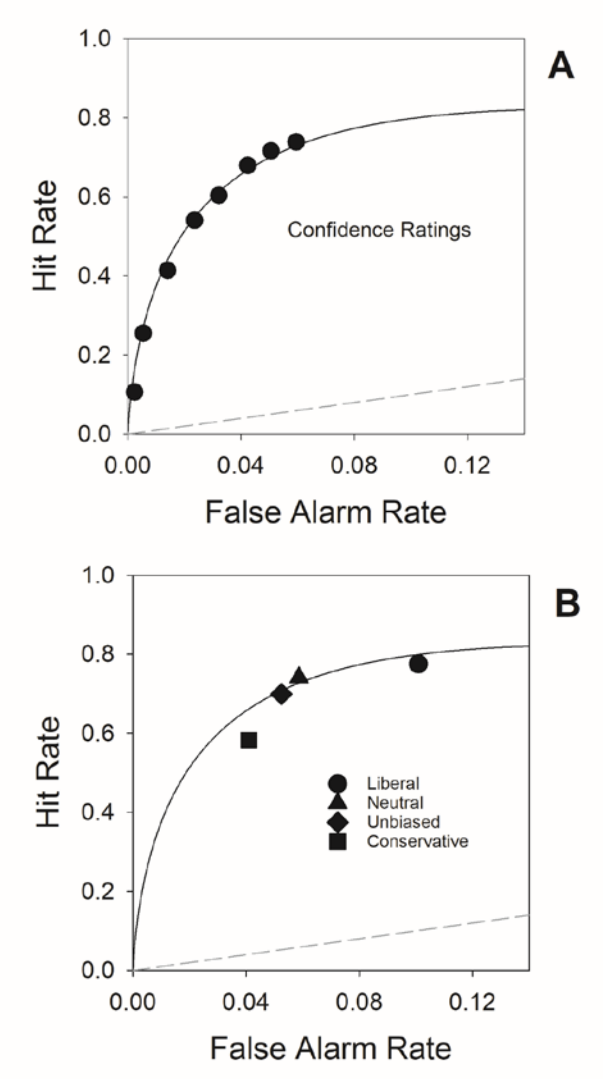


Figure 3. Simulated ROC data for the simultaneous and sequential lineup procedure as shown in Rotello and Chen (2016).

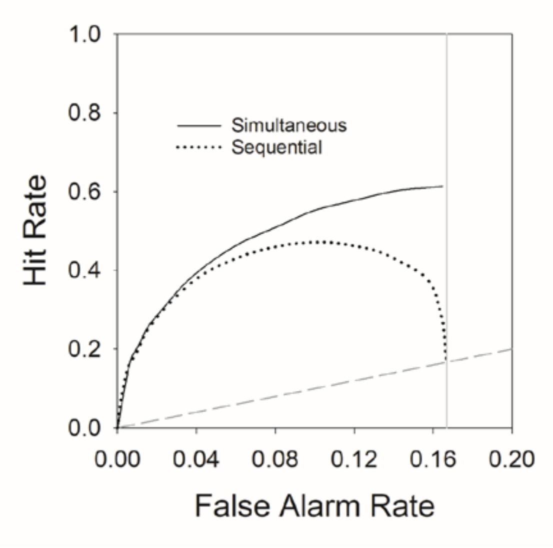


Figure 4. (A) Simple signal detection model with three different criteria used to simulate the ROC curves shown in B and C. (B) The binary ROC data that result from this simulation. (C) The confidence-based ROC data that result from this simulation.

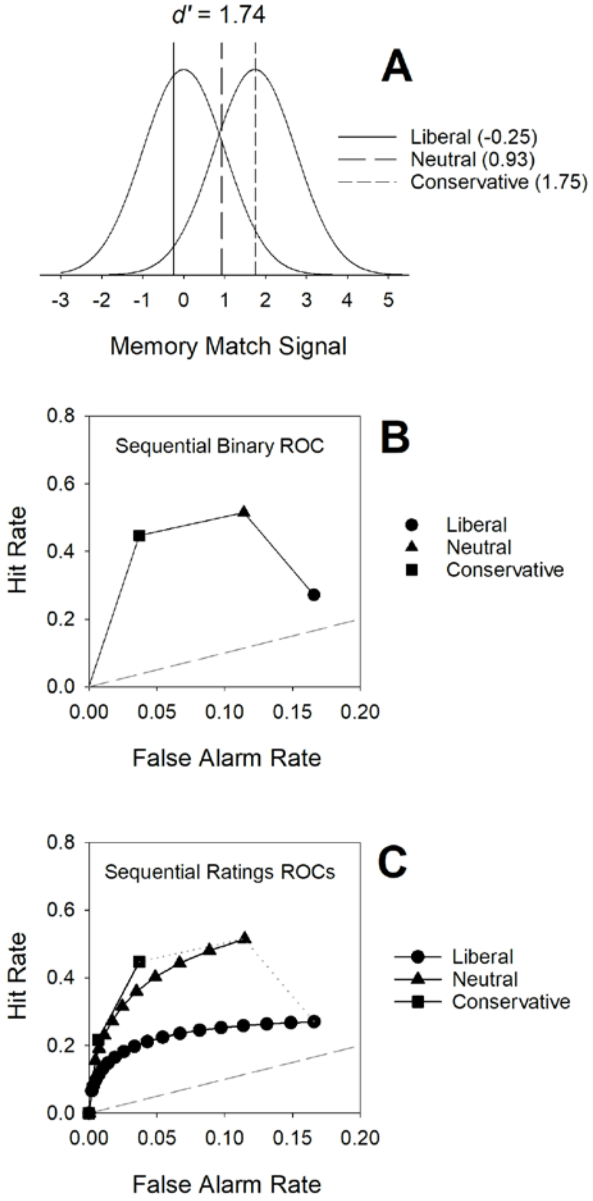


Figure 5. Confidence scale ranging from -100 (sure the face was not seen) to +100 (sure this is the face of the perpetrator). The overall criterion for counting an ID, thereby canceling any later IDs that might occur in the lineup, can be set to a liberal, neutral or conservative point on the confidence scale.

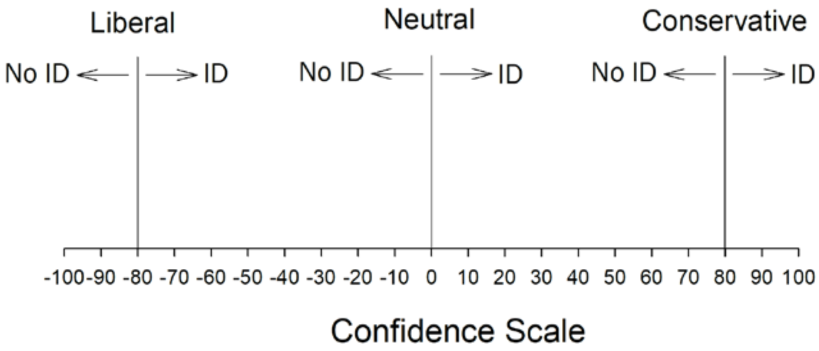


Figure 6. Simulated ROC data when the suspect appears in each of the six positions in the lineup when responding is neutral (A), liberal (B), or conservative (C).

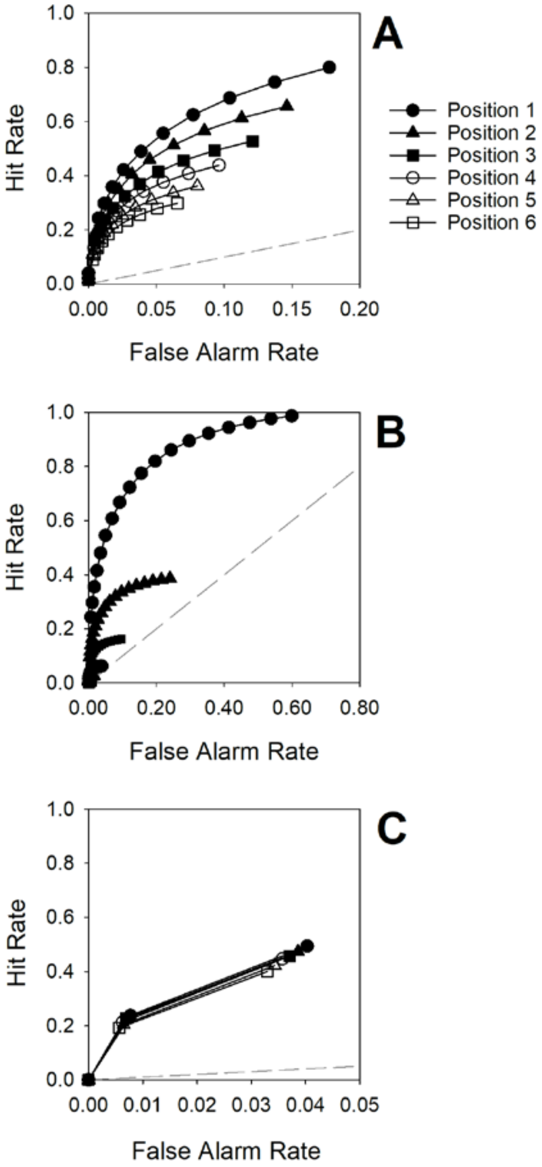


Figure 7. (A) The empirical binary ROC curves for the results of Experiment 1. (B) The empirical confidence-based ROC curves for the results of Experiment 1

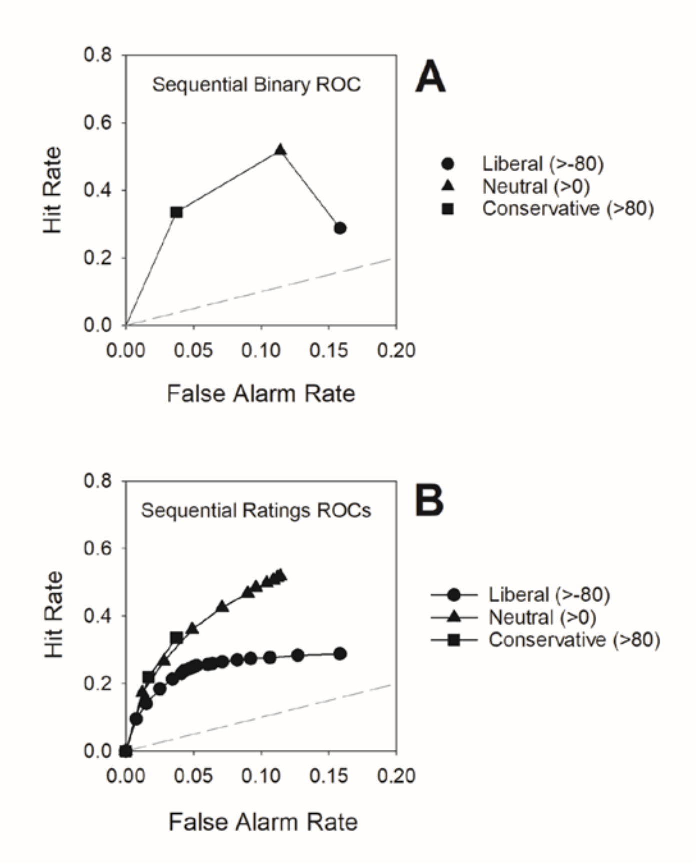


Figure 8. Empirical ROC curves from Experiment 1 when the suspect appears in each of the six positions in the lineup when responding is neutral (A), liberal (B), or conservative (C).

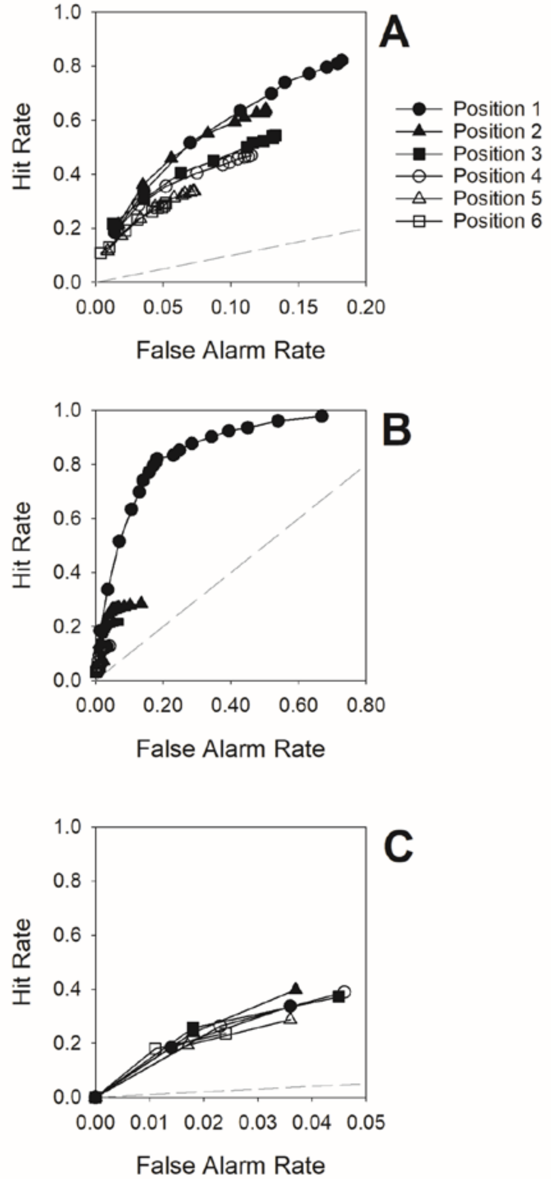


Figure 9. ROC curve for position 1 plotted against the ROC curve for positions 2-6 when only considering the subset of participants who have not yet made an identification prior to getting to each position.

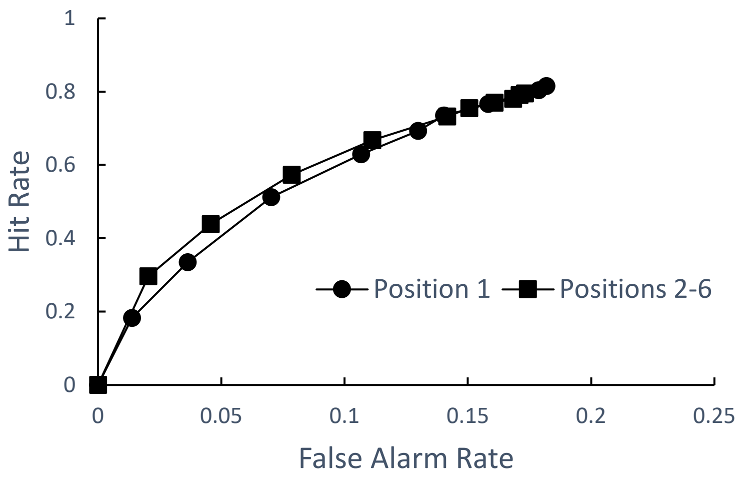


Figure 10. pAUC estimates for positions 1 through 6 when considering only the subset of participants at each position who have not yet made a prior identification. The solid horizontal line represents the mean value for the pAUC estimates for positions 2 through 6 (filled symbols), and the dashed horizontal lines represent the 95% confidence interval for those points. The pAUC estimate for positon 1 is shown as an open symbol.

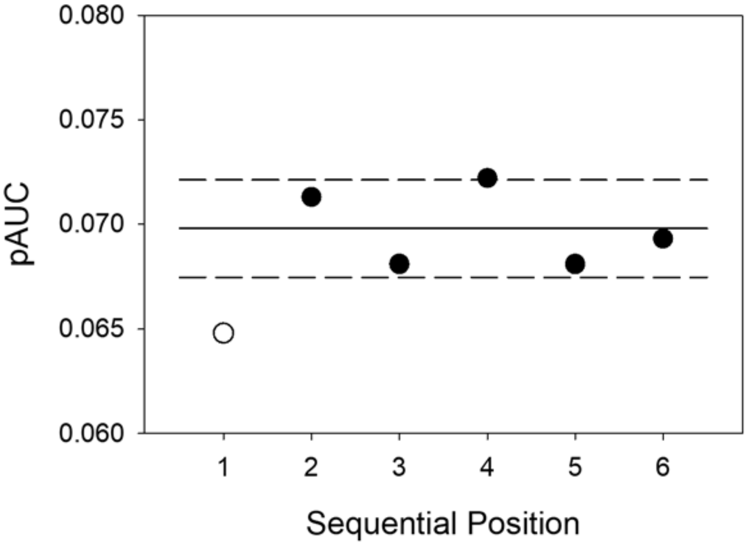


Figure 11. Median response times for the first "yes" decision to the suspect in target-present lineups (panel A) and to a filler in target-absent lineups (panel B). The solid horizontal line represents the mean value for the median response times from positions 2 through 6 (filled symbols), and the dashed horizontal lines represent the 95% confidence interval for those points. The median response time for positon 1 is shown as an open symbol.

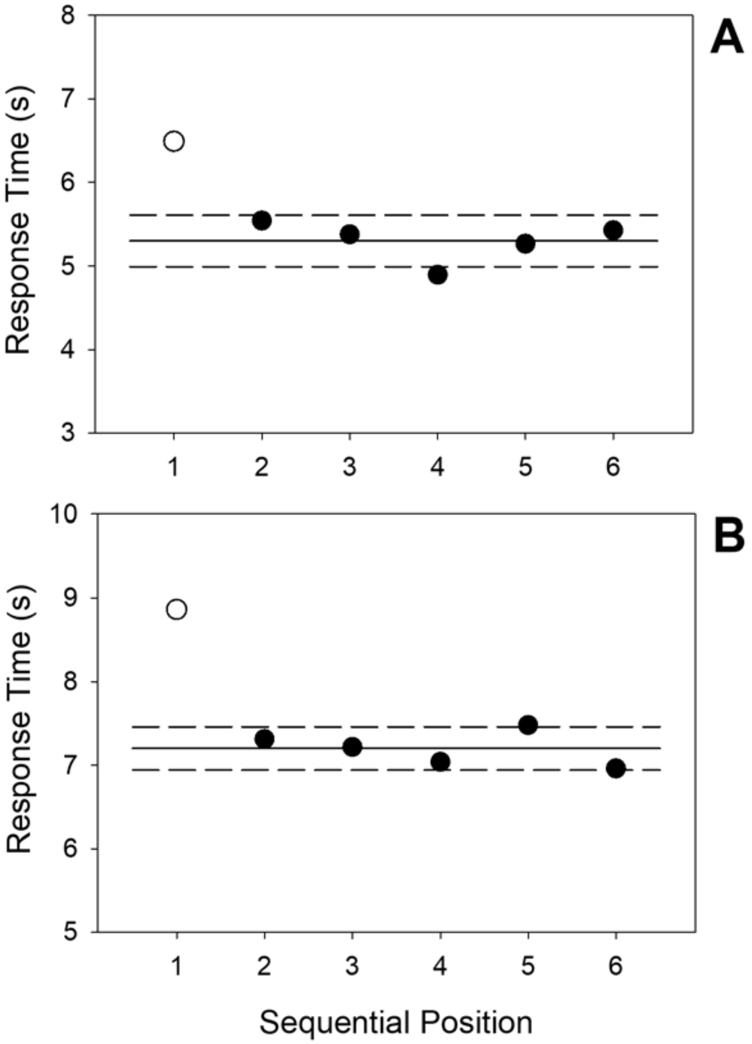


Figure 12. Empirical ROC curves for each position when the highest-confidence response is counted as an ID rather than the standard first-ID-only counting as an ID.

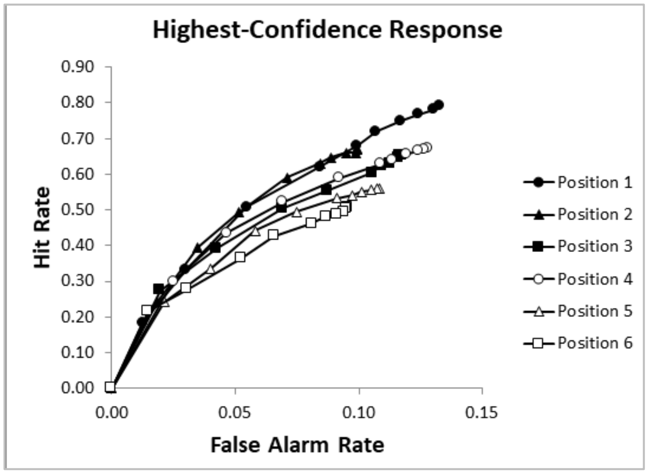


Figure 13. Empirical ROC curves for the showup vs. the sequential lineup in Experiment 2 (collapsed across positions).

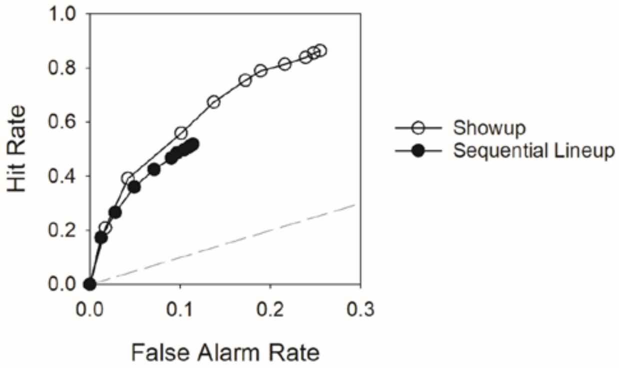


Figure 14. (A) The empirical binary ROC curves for the results of Experiment 2. (B) The empirical confidence-based ROC curves for the results of Experiment 2.

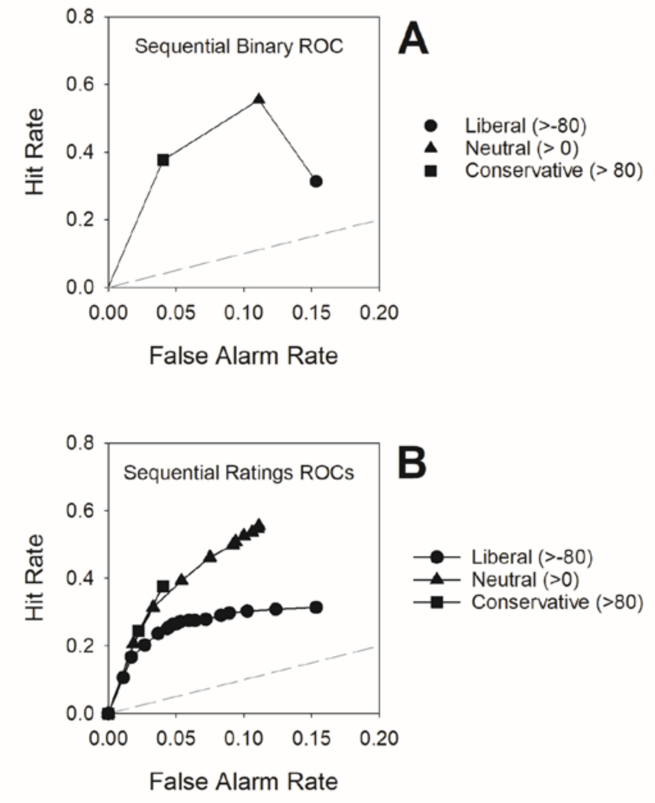


Figure 15. Empirical ROC curve for the showup and the sequential lineup for three different criteria.

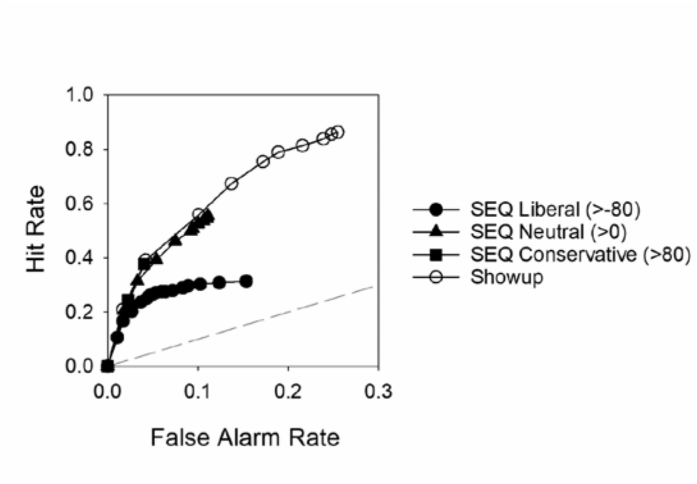
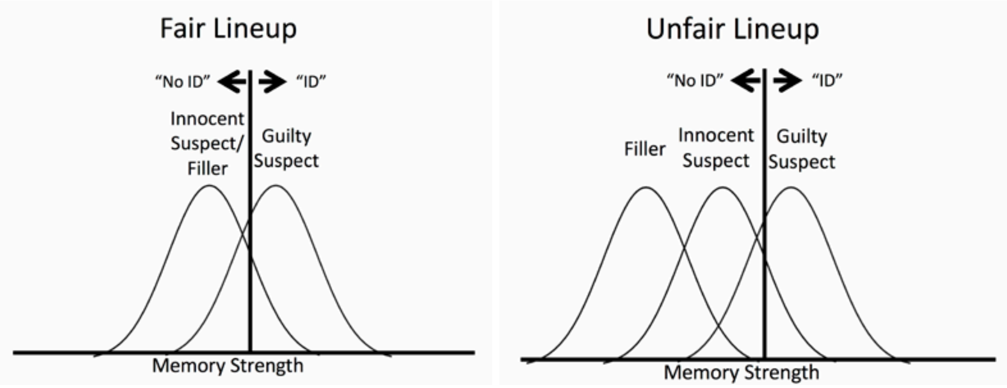


Figure 16. Distributions of memory strength values for fair and unfair lineups. With a fair lineup, the innocent suspect and filler distributions are the same. With an unfair lineup, however, the filler distribution has a lower average memory strength than the innocent suspect distribution.


Fair Lineup
“No ID”
“ID”
Innocent Suspect/ Filler
Guilty Suspect
Memory Strength
Unfair Lineup
“No ID”
“ID”
Filler
Innocent Suspect
Guilty Suspect
Memory Strength