A Signal-Detection-Based Diagnostic-Feature Model of Eyewitness Identification

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

[1]University of California, San Diego

[2]Royal Holloway, University of London

Author Note

John T. Wixted, Department of Psychology, University of California, San Diego. Laura Mickes, Department of Psychology, Royal Holloway, University of London.

This work was supported by the National Science Foundation SES-1155248 to John T. Wixted and Laura Mickes. The content is solely the responsibility of the authors and does not necessarily reflect the views of the National Science Foundation.

Correspondence concerning this article should be addressed to John Wixted (jwixted@ucsd.edu) or Laura Mickes (Laura.Mickes@rhul.ac.uk).

Abstract

The theoretical understanding of eyewitness identifications made from a police lineup has long been guided by the distinction between *absolute* and *relative* decision strategies. In addition, the accuracy of identifications associated with different eyewitness memory procedures has long been evaluated using measures like the diagnosticity ratio (the correct identification rate divided by the false identification rate). Framed in terms of signal-detection theory, both the absolute/relative distinction and the diagnosticity ratio are mainly relevant to response bias while remaining silent about the key issue of diagnostic accuracy, or discriminability (i.e., the ability to tell the difference between innocent and guilty suspects in a lineup). Here, we propose a signal-detection model of eyewitness identification, one that encourages the use of (and helps to conceptualize) receiver operating characteristic (ROC) analysis to measure discriminability. Recent ROC analyses indicate that the simultaneous presentation of faces in a lineup yields higher discriminability than the presentation of faces in isolation, and we propose a diagnostic feature-detection hypothesis to account for that result. According to this hypothesis, the simultaneous presentation of faces allows the eyewitness to appreciate that certain facial features (namely, those that are shared by everyone in the lineup) are non-diagnostic of guilt. To the extent that those non-diagnostic features are discounted in favor of potentially more diagnostic features, the ability to discriminate innocent from guilty suspects will be enhanced.

Keywords: Eyewitness Memory, Confidence and Accuracy, ROC Analysis, Signal-Detection Theory, Showups

A Signal-Detection-Based Diagnostic-Feature Model of Eyewitness Identification

Ever since Egan (1958) introduced signal detection theory to the field of memory, the understanding of how recognition decisions are made using standard laboratory tasks, such as memory for a recently presented list of words, has been guided by the distinction between *discriminability* (the ability to distinguish between targets that appeared on the list vs. foils that did not) and *response bias* (the overall tendency to classify a test item as being a target). By contrast, the understanding of how recognition decisions are made using more ecologically valid tasks related to eyewitness identification, such as memory for a perpetrator's face in a lineup, has been guided by the distinction between *absolute* vs. *relative* decision strategies (Wells, 1984). Viewed in terms of signal detection theory, a case can be made that this influential decision-strategy theory applies only to the question of why different eyewitness identification procedures give rise to different levels of response bias. But which procedure maximizes the ability of an eyewitness to tell the difference between innocent and guilty suspects presented in a lineup? That is, which procedure maximizes discriminability? Very little work has addressed that issue, and the reason may be that theories of discriminability in the domain of eyewitness identification are virtually nonexistent. The theory we propose below is intended to fill that gap, and it consists of two parts: (1) a general signal-detection-based model of eyewitness identification, and (2) a specific diagnostic feature-detection hypothesis that accounts for why the simultaneous presentation of faces in a lineup yields higher discriminability than alternative formats involving the presentation of faces in isolation. Before describing our new account, we first review the origins of the theory that has guided thinking about eyewitness identification more than any other theory over the last 30 years.

Absolute vs. Relative Decision Strategies

The distinction between absolute and relative decision strategies came about as part of a research-based effort to decrease the frequency of eyewitness misidentifications, which account for a high percentage of wrongful convictions that were later overturned by DNA evidence (Innocence Project, 2013). Two procedures that have long been used by the police to test an eyewitness's ability to recognize a perpetrator are the 1-person showup and the 6-person simultaneous lineup, both of which are thought to yield an uncomfortably high level of eyewitness misidentifications. A showup is simply an old/new recognition test in which a suspect (or a photograph of the suspect) is presented to an eyewitness who is then asked whether or not this individual is the perpetrator. A simultaneous lineup, by contrast, involves the side-by-side presentation of multiple individuals (or multiple photographs) to an eyewitness who is then asked if the perpetrator is present in the lineup (and, if so, to identify that individual). A typical lineup consists of the simultaneous presentation of six people, one of whom is the suspect (guilty or innocent) and five of whom are foils who are known to be innocent.

In an effort to reduce eyewitness misidentifications, several changes in the way that police conduct lineups have been proposed (see Clark, 2012, for a recent discussion of these reforms). One proposed change is to present the lineup photos one at a time (i.e., in isolation instead of simultaneously) for separate yes/no decisions, with the first "yes" decision being the only one that counts (Lindsay, 1999; Lindsay & Wells, 1985; Steblay, Dysart, Fulero, & Lindsay, 2001). This procedure, which is essentially a serial showup procedure, is known as the sequential lineup.

The police do not know if their suspect is innocent or guilty, so it is not possible to say whether a suspect who is identified from a police lineup by an eyewitness is a correct identification (a correct ID) or a false identification (a false ID). However, correct and false IDs

can be determined in forensically-relevant laboratory studies in which participants view a staged crime because, in that case, it is known whether or not the participant saw the suspect during an earlier study period. In a seminal laboratory study comparing simultaneous vs. sequential lineups, Lindsay and Wells (1985) reported that the false ID rate was much lower for sequential lineups than for simultaneous lineups (.17 vs. .43, respectively), but the correct ID rates were comparable and did not differ significantly for the two procedures (.50 vs. .58, respectively). Later meta-analyses revised that initial message, reporting that sequential lineups significantly reduce both the false ID rate and the correct ID rate (Steblay et al. 2001; Steblay, Dysart, & Wells, 2011). Thus, sequential lineups have both positive and negative effects (Clark, 2012). Despite the mixed pattern of results, the argument is often made that sequential lineups are superior to simultaneous lineups because, of the two procedures, sequential lineups typically yield a higher *diagnosticity ratio,* which is defined as the correct ID rate divided by the false ID rate (Steblay et al., 2011). The higher diagnosticity ratio associated with sequential lineups (which is often but not always found; see, for example, Gronlund, Carlson, Dailey, & Goodsell, 2009) means that a suspect identified from a sequential lineup is more likely to be guilty than a suspect identified from a simultaneous lineup. The combination of a lower false ID rate and a higher diagnosticity ratio seems to make a strong case in favor of the sequential lineup procedure.

From a theoretical standpoint, what explains the lower correct and false ID rates and higher diagnosticity ratios associated with sequential lineups? Wells (1984) proposed that simultaneous lineups create a tendency to identify the lineup member who most resembles the eyewitness's memory of the perpetrator. In the extreme case, the use of this *relative decision strategy* would lead eyewitnesses to always choose someone from a lineup, which is problematic in the case of target-absent lineups because of the high number of false identifications that would

occur. An alternative strategy is an *absolute decision strategy*, which Lindsay and Wells (1985) argued can be promoted by using a sequential lineup. Instead of choosing the lineup member who looks most like the perpetrator, eyewitnesses who use this strategy would evaluate each lineup member against an absolute decision criterion. If no one in the lineup yielded a strong enough match to the eyewitness's recollection of the perpetrator (i.e., a strong enough match to exceed the decision criterion), the lineup would be rejected. Because the tendency to identify someone from a lineup is lower when an absolute strategy is used, the likelihood of misidentifying an innocent suspect (i.e., the false ID rate) is correspondingly lower.

Although the use of an absolute strategy instead of a relative strategy explains why sequential lineups reduce the false ID rate compared to simultaneous lineups, there is nothing about the absolute/relative distinction that sheds theoretical light on why the diagnosticity ratio might also be higher for the sequential procedure. Yet, empirically, the diagnosticity ratio is often higher for sequential lineups, and that consistent empirical pattern of results led to the assumption that an absolute decision strategy not only decreases the overall tendency to identify someone from the lineup (reducing both the correct and false ID rates) but, for reasons unknown, also increases the diagnosticity ratio (Clark, Erickson, & Breneman, 2011).

Showups, even more so than sequential lineups, must promote the use of an absolute decision strategy because there are no other faces involved in the memory test that would allow a relative strategy to be used. Thus, it would be reasonable to suppose that showups also reduce the tendency to make a positive identification and also increase the diagnosticity ratio compared to simultaneous lineups. However, Lindsay and Wells (1985) argued that other factors associated with showups pull in the opposite direction (i.e., increasing the overall tendency to make a positive identification). In particular, showups are thought to be suggestive of guilt because only

one face is presented by the police to the eyewitness. By contrast, when a sequential lineup is used, the eyewitness is aware that more faces will be shown and is therefore protected from the suggestive nature of a 1-person showup.

In agreement with this line of reasoning, showups have not been found to decrease the likelihood of making an identification of a suspect compared to a simultaneous lineup (Clark, 2012; Clark & Godfrey, 2009). If anything, overall suspect identifications (correct ID rate + false ID rate) are higher for showups than for simultaneous lineups. Nevertheless, showups must involve the use of an absolute decision strategy. If the use of an absolute decision strategy yields a higher diagnosticity ratio than a relative strategy, then one might expect to find that showups, despite the higher suspect identification rates they induce, nevertheless yield a higher diagnosticity ratio than simultaneous lineups. However, this does not appear to be the case. Instead, showups appear to yield a *lower* diagnosticity ratio than simultaneous lineups (Clark & Godfrey, 2009, Table 7), which implies that an absolute decision strategy does not automatically translate into a higher diagnosticity ratio. Across the three eyewitness memory procedures (showup, simultaneous lineup, sequential lineup), the likelihood of making a suspect identification (highest for the showup and lowest for the sequential lineup) appears to be inversely related to the diagnosticity ratio (lowest for the showup and, perhaps, highest for the sequential lineup). This may not be a coincidence.

A Signal-Detection-Based Model of Eyewitness Identification

Signal-detection theory provides a useful alternative perspective for conceptualizing eyewitness identification because it clearly distinguishes discriminability from response bias (Clark et al., 2011; Duncan, 2006; Ebbesen & Flowe, 2002; Palmer & Brewer, 2012). In so doing, it sheds theoretical light on why the diagnosticity ratio behaves as it does, and it brings out

the fact that the theory of absolute vs. relative decisions can be construed as a theory of response bias – one that is silent about the arguably more important issue of discriminability.

A Signal-Detection Model for Showups

The Unequal-Variance Signal-Detection (UVSD) model has been widely used to conceptualize old/new recognition memory decisions since it was proposed more than a half-century ago (Egan, 1958). Because a showup is simply an old/new recognition test, the UVSD model (usually tested using lists of words) is a natural candidate for understanding eyewitness identification for faces tested in this manner.

In a typical list-memory study conducted in a cognitive psychology lab, the UVSD model is applied to data from a single participant who first studies a list of items and then completes a recognition test involving many targets and lures. In a forensically-relevant study, by contrast, the data reflect the performance of a group of participants who each watch a staged crime and then make only one decision about a target or a lure (depending on which is presented). Signal-detection-based analyses of group data do not characterize the performance of any individual participant, but this is also true of the correct and false ID rates computed from every forensically-relevant study. The goal of a forensically-relevant study is to characterize the population of participants who might be tested using a single-trial showup or a single-trial lineup (which is how memory is usually tested in police investigations). Although special considerations can arise when signal-detection analyses are performed on group data, standard signal-detection logic generally applies (Macmillan & Kaplan, 1985)[1].

In the context of eyewitness memory, the UVSD model specifies how face memory strength is distributed across guilty suspects (targets) and innocent suspects (lures). According to this account (illustrated in Figure 1), the mean and standard deviation of the target distribution

are greater than the corresponding values for the lure distribution (i.e., $\mu_{target} > \mu_{lure}$ and $\sigma_{target} > \sigma_{lure}$, respectively). Gaussian target and lure distributions are usually assumed, but the signal detection logic presented below applies equally to many other distributions (e.g., logistic, Weibull, lognormal, etc.). A key assumption of signal detection theory is that a decision criterion is placed somewhere on the memory strength axis, and an identification of the suspect is made if the memory strength of a face (target or lure) exceeds it. The correct ID rate is represented by the proportion of the target distribution that falls to the right of the decision criterion, and the incorrect ID rate is represented by the proportion of the lure distribution that falls to the right of the decision criterion. Figure 1 shows three possible placements of the decision criterion (liberal, neutral and conservative).

Discriminability vs. Response Bias. In signal-detection theory, discriminability is represented by the *degree of overlap* between the target and lure distributions. The more the memory signals associated with targets and lures overlap, the less able participants are to discriminate between the targets and lures on the recognition test. Discriminability is the same for the three signal detection models shown in Figure 1. Response bias, on the other hand, refers to the placement of the decision criterion for making a positive identification. Its placement is under the control of the participant, so it can be manipulated using pre-test instructions. For example, instructions that emphasize the importance of identifying a guilty suspect even if one's certainty is low will result in a relatively liberal setting of the criterion (illustrated by the model shown at the top left in Figure 1). A liberal criterion results in relatively high correct and false ID rates. Instructions that instead encourage an equal balance between falsely identifying innocent suspects and failing to identify guilty suspects will result in a more neutral response bias (illustrated by the model shown at the middle left in Figure 1), resulting in somewhat lower

correct and false ID rates. Instructions that encourage a high degree of certainty before making a positive identification will result in the use of a more conservative setting of the criterion (illustrated by the model shown at the bottom left in Figure 1), lowering both the correct and false ID rates even further.

The 3 pairs of correct and false ID rates from the different instructional conditions could be plotted against each other, thereby creating the receiver operating characteristic (ROC) shown to the right in Figure 1. Although the three points correspond to a single level of discriminability (which is why they fall on a single ROC curve), they would nevertheless be associated with three different diagnosticity ratios, with the highest diagnosticity ratio being associated with the most conservative criterion (point C on the ROC in Figure 1). The fact that the diagnosticity ratio increases monotonically as responding becomes more conservative corresponds to what has been found in several empirical studies (Gronlund et al., 2012; Mickes, Flowe, & Wixted, 2012), but no theoretical explanation for that phenomenon has yet been offered. However, signal-detection theory provides a natural explanation for why that effect occurs.

Signal Detection Theory and the Diagnosticity Ratio. An intuitive understanding of why signal-detection theory predicts that the diagnosticity ratio will increase as responding becomes more conservative is provided by the three signal-detection models shown in Figure 1. When responding is relatively liberal, the correct and false ID rates will both be high. In this case, the correct ID rate is 0.66, and the false ID rate is 0.34. Thus, the diagnosticity ratio is 0.66/0.34 = 1.94. If the criterion were moved even further in the liberal direction, very far to the left, the correct and false ID rates would both approach 1 (because 100% of both distributions would fall to the right of the criterion), and the diagnosticity ratio would approach 1 as well. A more interesting question is what the theory predicts when the criterion is instead moved in a more

conservative direction. Under those conditions, the false alarm rate (i.e., the proportion of the lure distribution to the right of the criterion) drops off more rapidly than the hit rate (i.e., the proportion of the target distribution to the right of the criterion). In the neutral condition illustrated in Figure 1, the correct ID rate is 0.47, and the false ID rate is 0.14, so the diagnosticity ratio has increased to 0.47/0.14 = 3.36. In the conservative condition, the criterion is so far to the right that the false ID rate is close to 0 (0.02 in this example) even though the correct ID rate remains well above zero (0.25 in this example). Moving the criterion to this conservative point on the memory strength axis increases the diagnosticity ratio to 0.25 / 0.02, or 12.5. Thus, as explained in more detail in the appendix, the basic tenets of signal-detection theory provide a theoretical interpretation of why the diagnosticity ratio steadily increases as responding becomes more conservative (even when discriminability remains unchanged).

A Signal-Detection Model for Simultaneous Lineups

The extension of the UVSD model to the simultaneous lineup is straightforward, up to a point. Consider first the case in which lineups are constructed in such a way that the innocent suspect (referred to as a "lure") does not resemble the perpetrator any more than the non-suspects (referred to as "foils") do. In other words, consider a fair lineup. In that case, the lure distribution and the foil distribution are one and the same. Conceptually, a target-present lineup with 6 members is represented by 5 random draws from the lure/foil distribution and one random draw from the target distribution, whereas the target-absent lineup is represented by 6 random draws from the lure/foil distribution.

If the innocent suspect *does* resemble the perpetrator more than the other 5 foils do (i.e., if an unfair lineup is used), then a different model involving a third distribution applies. In this case, an extra foil distribution is situated on the far left, a target distribution is situated on the far right,

and a lure distribution is situated somewhere in between. For an unfair lineup, the target-present lineup can be conceptualized as 5 random draws from the foil distribution and 1 random draw from the target distribution, and the target-absent lineup can be conceptualized as 5 random draws from the foil distribution and 1 random draw from the lure distribution. Whether the lineup is fair or unfair, the discriminability of interest concerns the overlap between the target and lure distributions (Figure 1).

The ideas presented above reflect the simplest possible extension of the UVSD model to the simultaneous lineup. However, there is still an additional step that is needed to complete the model, and that step is to identify the *decision strategy* used by the eyewitness. Clark et al. (2011) considered a variety of possible decision strategies that the eyewitness might use, one of which is as follows: first, identify the lineup member who best matches one's memory of the perpetrator (a relative strategy), and then identify that lineup member if the degree of match exceeds a decision criterion (an absolute strategy). If the degree of match does not exceed a decision criterion, the lineup is rejected. Clark et al. referred to this as the *Best Above Criterion* decision strategy. A variety of other strategies can be identified, and it is not known which strategy is actually used when participants identify someone from a simultaneous lineup. However, using simulations, Clark et al. found very small differences between the ROCs predicted by the various decision strategies they considered. Thus, the conceptual guidance provided by signal-detection theory remains the same across a variety of reasonable decision strategies.

As with showups, instructions can be used to manipulate the criterion used for identifications made from a lineup to generate an ROC. A more straightforward way to construct an ROC (for both showups and lineups) is to use confidence ratings. Instead of actually manipulating the decision criterion using instructions, one simply changes the decision rule used

to count a nominal positive ID as being an admissible positive ID (so to speak). We consider these issues in more detail next.

Confidence Ratings and Decision Criteria

Signal-detection theory provides a useful way to conceptualize the relationship between different levels of confidence and the corresponding correct and false ID rates. It is important to have a theory of how confidence ratings are made because the level of certainty expressed by an eyewitness when making a positive identification has a strong influence on jurors (e.g., Wells, Lindsay & Ferguson, 1979). In signal-detection theory, there is no fundamental distinction between a positive identification on the one hand and a confidence rating on the other because both are based on a decision criterion placed on the memory strength axis (Egan, 1958; Macmillan & Creelman, 2005). Figure 2 shows how signal-detection theory conceptualizes confidence ratings associated with positive IDs that are made using a 5-point scale, where 1 = Very Unsure and 5 = Very Sure for two different eyewitness identification procedures. Theoretically, the decision to identify a target or a lure is made when memory strength is sufficient to support a confidence rating of at least 1. Similarly, a decision to identify a target or a lure with the next highest level of confidence is made when memory strength is sufficient to support a confidence rating of at least 2 (and so on).

A non-obvious implication of this account is that more conservative responding can be achieved in two different ways. First, in the absence of confidence ratings (and as noted earlier), instructions can be used to encourage participants to be more conservative about making an identification (i.e., to not make an identification unless they are quite confident of being correct). Second, if confidence ratings are collected for all positive IDs, more conservative responding can be achieved by only counting identifications that are made with relatively high confidence (e.g., a

rating of 4 or more on a 5-point scale) while treating the lower confidence identifications as non-identifications. According to signal detection theory, these two biasing strategies achieve the same result, namely, a more conservative decision criterion. Moreover, it follows that ROC analysis can be performed from confidence rating data without having to conduct multiple conditions involving different biasing instructions (Mickes et al., 2012; Gronlund, Wixted, & Mickes, in press). To do so, one pair of correct and false ID rates can be computed using only those positive IDs that were made with a rating of 5, a second pair can be computed using only those positive IDs that were made with a rating of 4 or 5, and so on until, finally, a pair of correct and false ID rates are computed counting all positive IDs (i.e., IDs associated with confidence ratings of 1, 2, 3, 4 or 5). This final correct and false ID rate pair corresponds to what is usually reported and analyzed in studies of eyewitness memory (e.g., Lindsay & Wells, 1985). However, each one of the correct and false ID rate pairs that make up the ROC are equally relevant.

The connection between ROC data and the signal-detection interpretation of discriminability is also illustrated in Figure 2. Procedure A might be a (fair) simultaneous lineup and Procedure B might be a showup. In this hypothetical example, Procedure A results in a larger separation of the innocent and guilty suspect distributions than Procedure B. The corresponding confidence-based ROC data shown to the right in Figure 2 reflect those differences in discriminability. That is, the hypothetical confidence-based ROC associated with Procedure A bows further above from the diagonal line of chance performance than the corresponding ROC for Procedure B. In practice, one proceeds in the other direction, from data (i.e., the ROC) to theoretical interpretation. Thus, the lineup procedure that yields the highest empirical ROC is the one that theoretically best facilitates the discrimination between innocent and guilty suspects by reducing the overlap between the corresponding memory strength distributions.

With regard to real world criminal investigations, the considerations discussed above pertain to the confidence expressed by an eyewitness at the time of the initial identification because confidence ratings that are made over the course of a lengthy police investigation can be easily contaminated (Wells, Small, Penrod, Malpass, Fulero & Brimacombe, 1998)[2]. For example, repeated identifications (e.g., across multiple lineups) can increase the familiarity of the suspect's face, inappropriately inflating confidence on later tests (e.g., Godfrey & Clark, 2010). In addition, confirming feedback provided to the eyewitness from the lineup administer can lead to inflated estimates of confidence when witnesses are later asked to retrospectively assess how confident they were at the time of their initial identification (Wells & Quinlivan, 2009). For these and other reasons, a Department of Justice panel recommended that eyewitness confidence should be assessed and recorded when the identification is first made (Technical Working Group for Eyewitness Evidence, 1999). These guidelines stipulate that the investigator should "Record both identification and nonidentification results in writing, including the witness' own words regarding how sure he/she is" and should "Ensure that no materials indicating previous identification results are visible to the witness" (p. 38). Elsewhere, the guidelines caution that "If an identification is made, avoid reporting to the witness any information regarding the individual he/she has selected prior to obtaining the witness' statement of certainty" (p. 33). When obtained under the conditions recommended by the Department of Justice guidelines, confidence ratings are potentially very useful, and they are intimately related to the criterion-shift issues discussed above. For example, a court of law that takes under advisement only those suspect identifications that were made with a moderately high degree of confidence (or more) has decided not to focus on decisions that correspond to the rightmost ROC point but to focus instead on decisions that correspond to an ROC point to the left of that.

Absolute vs. Relative Decision Strategies and Response Bias

With the signal-detection interpretation of eyewitness identification in mind, we now consider what aspect of performance the theory of absolute vs. relative decision strategies addresses. Our suggestion is that it is a theory about response bias (i.e., the tendency to choose), not a theory about discriminability. Indeed, in his original paper on the topic, Wells (1984) wrote "It is possible to construe of the relative judgments process as one that yields a response bias, specifically a bias to choose someone from the lineup" (p. 94). A witness who relies on the relative judgment strategy, which is theoretically promoted by simultaneous lineups, has a tendency to choose the lineup member who looks most like the perpetrator. A witness who relies on the absolute judgment strategy, which is theoretically promoted by sequential lineups, has a lesser tendency to choose a lineup member because faces presented to the eyewitness in isolation are compared against an absolute standard. But an absolute standard is nothing more than a decision criterion. Moreover, because not every witness presented with a simultaneous lineup chooses someone, an absolute standard (i.e., a decision criterion) must be used for simultaneous lineups as well. The difference is that, in the sequential lineup, a more conservative standard is theoretically used (reflecting the lesser tendency to choose) compared to a simultaneous lineup.

It is possible to conceptualize absolute vs. relative comparisons in such a way that discriminability, not bias, is predicted to be affected (e.g., Clark et al., 2011). Indeed, our own theory (presented in a later section) holds that the relative comparisons afforded by a simultaneous lineup enhance discriminability. However, our point here is that there is nothing in the original formulation of the theoretical distinction between absolute and relative decision strategies that makes any prediction about discriminability. Instead, the original formulation focuses solely on the *tendency to choose* someone from the lineup, and it provides an intuitively

understandable explanation for why that tendency is lower (i.e., for why a more conservative criterion is used) when the members of a lineup are presented sequentially rather than simultaneously.

If the sequential presentation of faces in a lineup had no effect other than to induce more conservative responding (thereby lowering the overall correct and false ID rates), then, according to signal detection theory, sequential lineups should also have a higher diagnosticity ratio than simultaneous lineups. A recent review of the literature concluded that sequential lineups have exactly that effect (Steblay et al., 2011). The overall pattern of results associated with switching from the simultaneous to the sequential procedure can be summarized in simple terms (e.g., for policymakers or for the general public) by saying that the sequential procedure substantially reduces mistaken identifications while reducing correct identifications to a lesser extent. When that pattern of results is considered in light of the well-known problems associated with eyewitness misidentifications, the superiority of the sequential procedure might seem self-evident. However, intuition notwithstanding, there is nothing inherently superior about more conservative responding. The optimal bias is not the most conservative bias; instead, the optimal bias (and, therefore, the optimal point on the ROC) is a joint function of subjective values associated with the different decision outcomes and the (unknown) base rate of guilty suspects in lineups (see Mickes et al., 2012, p. 373, for a discussion of these issues).

The overall correct and false ID rates associated with simultaneous and sequential lineups, which include positive IDs made with any level of confidence, represent the rightmost point of each procedure's ROC. The data reviewed by Steblay et al. (2011) suggest that rightmost point of the sequential ROC falls to the left of (i.e., it reflects more conservative responding than) the rightmost point of the simultaneous ROC. This finding may have limited forensic relevance

because courts of law generally attach little or no weight to positive IDs made with the lowest levels of confidence. As such, they are interested in points that fall more to the left on the ROC (i.e., points that exclude IDs made with the lowest levels of confidence, or that exclude IDs made with the lowest and next-to-lowest levels of confidence, and so on, depending on how far to the left one goes). It is not possible to say which point on the ROC best corresponds to the higher standard typically used in a court of law, but it seems certain that the rightmost point on the ROC (the one associated with the most liberal criterion for making a positive ID) is the least relevant. Instead of focusing solely on the rightmost point of an ROC, it would be more informative to examine the entire ROC associated with a particular eyewitness identification procedure. Somewhere along that ROC is what the court wants to know.

The very existence of an empirical ROC shows that response bias can be easily varied over a wide range using either lineup procedure. Thus, varying response bias does not pose a great challenge (even though determining the *optimal* response bias does). We suggest that a theory about response bias is less important than a theory about discriminability. The reason is that the lineup procedure that yields the highest discriminability can simultaneously achieve a higher correct ID rate and lower false ID rate than the alternative procedures. As such, it is the unambiguously superior procedure (unlike a procedure that merely induces more conservative responding with respect to the rightmost ROC point).

The discussion presented above implicitly assumed that the more conservative rightmost ROC point associated with the sequential procedure falls on the same curve as the more liberal rightmost ROC point associated with the simultaneous procedure. However, whether or not that is true is an empirical question, one that has only recently been addressed. It is possible that the rightmost point of the sequential ROC lies on a different curve than the rightmost point of the

simultaneous ROC, in which case the two procedures would differ not only in response bias but also in discriminability.

Recent Empirical ROC Analyses

The first three studies using ROC analysis have so far not found any evidence that the sequential procedure is diagnostically superior to the simultaneous procedure (Dobolyi & Dodson, in press; Gronlund et al., 2012; Mickes et al., 2012). Instead, all three studies unexpectedly found a significant advantage for the *simultaneous* procedure. Figure 3 shows data reported by Mickes et al. (2012), which used a forensically-relevant experimental design, showing that the simultaneous lineup yields higher discriminability than the sequential lineup. With regard to simultaneous lineups vs. showups, another recent ROC analysis found a significant advantage for the simultaneous procedure in that case as well (Gronlund et al., 2012). Whereas the former result comes as a surprise (because many have long assumed that the sequential procedure is diagnostically superior to the simultaneous procedure), the latter result does not. For example, Clark (2012) recently summarized the relevant evidence on showups as follows: "Averaging over 15 comparisons, lineups show lower false identification rates (.11) and slightly higher correct identification rates (.43) than showups (.18 and .41, respectively)" (p. 244). The fact that the simultaneous lineup is associated with both a lower false ID rate and a slightly higher correct ID rate than the showup makes it reasonably safe to conclude (even in the absence of ROC analysis) that the simultaneous lineup is diagnostically superior to the showup.

Together, these findings indicate that, for some as yet unexplored reason, presenting faces simultaneously enhances one's ability to tell the difference between innocent and guilty suspects compared to when they are presented in isolation (either in a showup or as part of a sequential lineup). In other words, presenting faces simultaneously enhances *discriminability*. Why would a

simultaneous lineup enhance one's ability to discriminate innocent suspects from guilty suspects compared to procedures that involve the presentation of faces in isolation? There is nothing in the signal-detection framework outlined above that answers that question, but we turn now to a theoretical principle that does.

Diagnostic Feature-Detection Hypothesis

Gibson (1969) reviewed a large body of evidence from the perceptual learning literature and identified a key principle that we suggest is relevant to the different levels of discriminability supported by different eyewitness memory procedures. She concluded that an important step in perceptual learning – that is, learning to discriminate similar objects – involves the detection of distinctive features. For example, to the non-expert, it is nearly impossible to tell the difference between an x-ray that shows evidence of a tumor vs. an x-ray that does not. One reason the task is so difficult is that, to the non-expert, it is not clear exactly which features of the x-ray to focus on and which to ignore in order to make the discrimination. To the extent that the non-expert focuses on non-diagnostic features of the x-ray (i.e., features that are common to x-rays that contain evidence of a tumor and x-rays that do not), discriminability will be impaired (Myles-Worsley, Johnston & Simons, 1988). In perceptual learning experiments, it has often been found that presenting similar objects simultaneously facilitates the detection of the distinctive features that serve to differentiate the objects (thereby enhancing discriminability between them) compared to when the objects are presented sequentially (Gibson, 1969).

Why would the simultaneous presentation of similar objects be advantageous compared to sequential presentation? Presumably, it is the opportunity for stimulus comparison that simultaneous presentation affords, and that opportunity facilitates learning to discriminate similar faces (not just similar x-rays). For example, Mundy, Honey, & Dwyer (2007) investigated the

ability of participants to discriminate between pairs of very similar faces that were created using a morphing program. They found that the simultaneous presentation of two similar faces on each trial resulted in better performance compared to when the similar faces were presented sequentially. Theoretically, the simultaneous presentation of two similar faces made it easier to identify their distinctive features (and to ignore their many common features) compared to when the faces were presented sequentially.

We propose that a similar process involving the detection of distinctive (and, therefore, diagnostic) features plays an important role at the time an eyewitness identification is made. Consider, for example, an eyewitness who sees a White male in his early 20s rob a liquor store. In addition to noticing the age, race and gender of the perpetrator, the eyewitness might also notice that the perpetrator has an oval face and small eyes[3]. Next imagine that the police identify a suspect who matches the general description provided by the eyewitness (namely, a White male in his early 20s) and that the police present that suspect to the eyewitness using either a showup or a 6-person simultaneous lineup. In the simultaneous lineup, all 6 members will match the general description of the suspect – that is, the suspect and the foils will all be young White males.

Consider four facial features that the eyewitness might attach weight to when trying to decide whether or not the suspect (or a foil) is the perpetrator. These four features are: age, race, shape of face, and size of eyes. The first two features (age and race) are non-diagnostic because they are shared by innocent and guilty suspects. These features are shared because they served as a basis for apprehending the suspect in the first place. In other words, any individual who did not possess those features would probably not be picked up by the police and presented to the eyewitness for possible identification (and the same consideration would apply to the choice of

foils as well). By contrast, the other two features (oval face and small eyes) are potentially diagnostic because, not having served as a basis for apprehending the suspect, they are less likely to be shared by innocent and guilty suspects.

When a face is presented in isolation in a showup or as part of a sequential lineup, there is no obvious indication to the eyewitness that some features are diagnostic and others are not. To the extent that the non-diagnostic features are given weight under those circumstances, the ability to discriminate innocent from guilty suspects will suffer. In a simultaneous lineup, by contrast, it is immediately apparent to the eyewitness that everyone in the lineup shares certain non-diagnostic features (e.g., it is immediately apparent that age and race are features that are of no use in deciding whether or not the perpetrator is in the lineup). For that reason, the eyewitness will be encouraged to attach weight to features that might be diagnostic while discounting features that are non-diagnostic. To the extent that they do, the ability to discriminate an innocent suspect from a guilty suspect will be enhanced. That is the essence of the diagnostic feature-detection hypothesis.

To formalize this simple example (and to illustrate how simultaneous lineups might enhance discriminability), assume that the memory strength of any feature that was not observed by the eyewitness (e.g., square face) has a mean of 0 ($\mu_{innocent} = 0$) and a standard deviation of 1 ($\sigma_{innocent} = 1$) across all possible innocent suspects. However, a feature that was observed by the eyewitness (e.g., oval face) has a mean of 1 ($\mu_{guilty} = 1$) and a standard deviation of 1.25 ($\sigma_{guilty} = 1.25$) across all possible suspects. These numbers correspond to the UVSD model as applied to a particular diagnostic feature (i.e., $\sigma_{innocent} / \sigma_{guilty} = 0.80$). Using this feature alone (e.g., oval face), the ability of a witness to discriminate an innocent suspect from a guilty suspect could be assessed by the d_a statistic, where:

$$d_a = \frac{\mu_{guilty} - \mu_{innocent}}{\sqrt{(\sigma^2_{guilty} + \sigma^2_{innocent})/2}}$$

d_a is much like d' except that it allows the two underlying distributions to have unequal variances. A d_a value of 0 would represent a complete inability to discriminate innocent from guilty suspects using that single feature (the corresponding ROC would fall on the diagonal line of chance performance), whereas larger values reflect increasingly accurate discriminability (corresponding to ROC data that increasingly bow away from the diagonal line of chance performance towards the upper left corner). In terms of signal detection theory, a d_a of 0 would be represented by two completely overlapping memory strength distributions for innocent and guilty suspects. By contrast, the larger d_a is, the lower the degree of overlap between those two distributions. When d_a is very large (e.g., 4 or more), the overlap between the two distributions would be negligible and accuracy would be nearly perfect. Using the hypothetical values presented above for a single diagnostic feature, d_a equals 0.88 for a decision based on one feature (oval face).

Presumably, an eyewitness does not rely on only one feature to decide whether or not the suspect is the perpetrator. Instead, multiple features are combined to make that decision. The upper half of Table 1 (Showup) presents hypothetical values for the 4 features under consideration here. The first two features are race (f1) and age (f2). These features were observed by the eyewitness, so their memory strength characteristics reflect the fact that these features are "old" (mean = 1, standard deviation = 1.25, which is to say that variance = 1.56). Note that these features are both non-diagnostic in the sense that the mean and standard deviation associated with them are shared by innocent and guilty suspects alike. Because these features were seen before, both innocent and guilty suspects (as well as foils who also share those features) may seem somewhat familiar to the eyewitness. The values for the next two features, face shape (f3) and

eye size (f4), are not shared by innocent and guilty suspects because the qualities of these features that were observed by the eyewitness (oval face and small eyes) were not used by the police to identify the suspect (or to choose foils for the simultaneous lineup). Because these features differ for innocent and guilty suspects, they are diagnostic of innocence or guilt. In Table 1, the mean and standard deviation for these features are 0 and 1, respectively, for innocent suspects (because these features, such as large eyes and a square face, are "new") and are 1 and 1.25, respectively, for guilty suspects (because these features, such as small eyes and an oval face, are "old").

The simplest way (but not the only way) to combine these feature values to arrive at an aggregate memory strength value is to sum them, yielding a random variable with a mean equal to sum of the component means and with a variance equal to the sum of the component variances (assuming independence across features). An eyewitness presented with a showup might be inclined to sum across all four features because it would not be immediately apparent that some features are diagnostic and some are not. Additively combining the diagnostic and non-diagnostic features to produce an aggregate memory strength variable for an innocent suspect yields the values shown in the rightmost column (∑) of Table 1. Specifically, $\mu_{innocent} = 2$ and $\sigma^2_{innocent} = 5.12$. Doing the same for a guilty suspect yields $\mu_{guilty} = 4$ and $\sigma^2_{guilty} = 6.24$. Using these values, the measure of discriminability would be $d_a = 0.84$. This value reflects the fact that both diagnostic and non-diagnostic features contributed to the decision.

An eyewitness presented with a 6-person simultaneous lineup would, according to our theory, be less inclined to incorporate the non-diagnostic facial features into the aggregate memory strength variable. The reason is that the eyewitness would realize that these features, which are obviously shared by everyone in the lineup, would not help to differentiate an innocent

suspect from a guilty suspect. Thus, the shared features would be given less weight. Attaching zero weight to those shared features yields the example shown in the lower half of Table 1. Summing across only the diagnostic features (f3 and f4) yields $\mu_{innocent} = 0$ and $\sigma^2_{innocent} = 2$ for the innocent suspect and $\mu_{guilty} = 2$ and $\sigma^2_{guilty} = 3.12$ for the guilty suspect. The measure of discriminability based on these values is $d_a = 1.25$. In other words, discriminability is higher for the simultaneous condition. In terms of signal-detection theory, a higher d_a means that the distribution representing innocent suspects (and foils) is more separated from the distribution representing guilty suspects. In terms of ROC analysis, a higher d_a means that the ROC curve obtained using a simultaneous lineup bows further away from the diagonal line of chance performance than the ROC obtained using a showup (a prediction that corresponds to a finding recently reported by Gronlund et al., 2012).

General Discussion

Our goal was to develop a signal-detection-based model of eyewitness identification that draws a clear distinction between discriminability and response bias. Our account suggests that the prevailing view of eyewitness identification based on the distinction between absolute vs. relative judgments is a theory about response bias. Although the absolute/relative distinction may explain why sequential lineups yield more conservative responding than simultaneous lineups, it does not speak to the more critical issue of discriminability. Signal detection theory illustrates the concept of discriminability, and it encourages the use of ROC analysis to measure it. The basic tenets of signal detection theory also explain why the diagnosticity ratio (a common measure of probative value) increases as responding becomes more conservative and, therefore, why that measure is unable to identify the better lineup procedure.

As in the field of medicine, where the goal is to identify the diagnostic test that best discriminates between the presence vs. absence of a disease (e.g., Metz, 2006; Swets, 1996), the goal of eyewitness memory researchers should be to identify the identification procedure that best discriminates between the presence vs. absence of a guilty suspect. The procedure that yields the highest discriminability is the one that, empirically, yields the highest ROC and that, theoretically, best separates the distribution of memory signals associated with innocent vs. guilty suspects (Figure 2). Recent ROC analyses indicate that the simultaneous lineup yields higher discriminability than both showups and sequential lineups (Dobolyi & Dodson, in press; Gronlund et al., 2012; Mickes et al., 2012).

How do these considerations relate to the field's longstanding goal of reducing eyewitness misidentifications? Sequential lineups reduce the overall false ID rate compared to both simultaneous lineups and showups, but, in comparison to simultaneous lineups, the price paid is a reduced correct ID rate and reduced discriminability. Can a reasonable case be made that the benefit of reduced false IDs associated with the use of sequential lineups is worth the cost? In our view, the answer is clearly no. If the goal is to achieve fewer false IDs despite the loss of correct IDs, then the solution would be to use a more conservative decision rule in conjunction with the diagnostically superior lineup procedure (i.e., the simultaneous procedure), not to switch to a diagnostically inferior lineup procedure to achieve more conservative responding. The reason is that the diagnostically superior procedure can achieve a lower false ID rate and, at the same time, a higher correct ID rate than the diagnostically inferior procedure.

Why does the simultaneous presentation of faces in a lineup enhance the ability of participants to tell the difference between innocent and guilty suspects (compared to showups and sequential lineups)? In addition to the signal-detection framework we proposed, we also proposed

a diagnostic feature-detection hypothesis to explain that result. This hypothesis holds that the memory signal associated with a face in a lineup is a function of multiple facial features (age, race, shape of face, size of eyes, etc.). Some of those features are diagnostic (i.e., they differ for innocent and guilty suspects) and some are not (i.e., they are the same for innocent and guilty suspects). An innocent suspect picked up by the police and the foils chosen by the police to fill out a lineup will presumably be chosen because they match the general description of the suspect provided by the eyewitness. The general description (e.g., young White male) is therefore a description of the *non-diagnostic features*. When a face is presented in isolation, either in a showup or as part of a sequential lineup, there is no cue indicating that some features are diagnostic and others are not. To the extent that the non-diagnostic features are given weight, the ability to discriminate innocent from guilty suspects will suffer (cf. Tversky, 1977). In a simultaneous lineup, by contrast, the presentation of 6 similar faces makes it clear to the eyewitness that all lineup members shares certain non-diagnostic features. Any eyewitness who notices those non-diagnostic (i.e., shared) features will presumably attach little or no weight to them and will instead focus on features that stand a better chance of being diagnostic. To the extent that the non-diagnostic features are discounted, the ability to discriminate an innocent suspect from a guilty suspect will be enhanced. This hypothesis explains why simultaneous lineups yield higher discriminability compared to showups and sequential lineups.

A sequential lineup eventually provides the same diagnostically useful information that is immediately apparent to the eyewitness when a simultaneous lineup is used. Thus, for example, by the end of the sequential lineup, the witness will presumably appreciate the fact that everyone under consideration matches the general description of the perpetrator (e.g., all are young White males). Thus, the diagnostic feature-detection hypothesis would have to predict that when

innocent and guilty suspects are placed later in the sequential lineup (e.g., in position 5), discriminability, measured using ROC analysis, should be higher than when innocent and guilty suspects are placed earlier in the sequential lineup (e.g., in position 2). As it happens, Gronlund et al. (2012) conducted precisely this experiment, and the results confirmed this prediction. When the innocent and guilty suspects always appeared in position 2 of a sequential lineup, the ROC was similar to that of a showup (well below that of the simultaneous lineup). When the innocent and guilty suspects always appeared in position 5 of a sequential lineup, the ROC was similar to (even slightly higher than) that of a simultaneous lineup. Theoretically, we propose that this result occurred because, by the time the participant reached position 5 without having yet chosen anyone, the non-diagnostic features were as clear to the participant as they would have been had the faces been shown simultaneously. Goodsell, Gronlund and Carlson (2010) considered the possibility that a witness might rely on more diagnostic features as a sequential lineup unfolds (just as we are suggesting). Our diagnostic feature-detection hypothesis goes beyond that by offering an explanation of how it is that witnesses become aware of what the more diagnostic features are during the course of sequential testing (namely, by taking note of common features, which are then discounted). Moreover, the same explanation that accounts for the sequential position effect also theoretically accounts for why simultaneous lineups yield higher discriminability than showups and sequential lineups.

A recent study found that accuracy was higher when a distinctive feature on the perpetrator (e.g., a scar) was replicated across everyone in a simultaneous lineup compared to when it was concealed on everyone in the lineup (Zarkadi, Wade & Stewart, 2009). On the surface, the diagnostic feature-detection hypothesis does not predict this result because when the salient feature appears on everyone in the lineup, it is not diagnostic and should be discounted.

Thus, because the relevant feature is missing in the concealment condition and is theoretically discounted in the replication condition, the two conditions should yield the same level of performance. However, the diagnostic feature-detection hypothesis has nothing to say about what happens when a witness clearly remembers a face that is now missing a permanent feature (something that would selectively occur in the concealment condition). Conceivably, if the participant has a clear memory of the perpetrator with a scar, but that scar is missing on the only target-present lineup member who strongly resembles the perpetrator, the participant might use a recall-to-reject strategy to conclude that the perpetrator is not in the lineup (and would reject the lineup on that basis). This would have the effect of turning what would otherwise be correct IDs on target-present trials (in the replication condition) into lineup rejections (in the concealment condition). As a result, the correct ID rate would be lower and the miss rate would be higher in the replication condition. On target-absent trials, by contrast, participants who have a clear memory of the perpetrator's face would already be inclined to reject the lineup whether or not the foils in the lineup contain the relevant feature. Thus, the false ID rates should be similar in the two conditions. This corresponds to the pattern of results reported by Zarkadi et al. Because of the availability of a recall-to-reject strategy in the concealment condition, these data do not provide a strong test of the diagnostic feature-detection hypothesis. By contrast, the sequential lineup position data reported by Gronlund et al. (2012) do provide a direct test of that hypothesis.

Conclusion

The theoretical understanding of eyewitness identification tested using showups, simultaneous lineups and sequential lineups has long been guided by the distinction between absolute and relative decisions (Wells, 1984). In addition, eyewitness memory procedures have long been evaluated using the diagnosticity ratio (or a closely related measure). The absolute-vs.-

relative judgment theory may explain why sequential lineups yield more conservative responding than simultaneous lineups, but it does not address whether one procedure yields higher discriminability than the other. To address that issue, we propose an alternative theoretical perspective based on signal-detection theory, which makes it clear why ROC analysis is the method needed to identify the most diagnostically accurate procedure and what a higher ROC theoretically means. We also proposed a novel theory of why the simultaneous presentation of faces yields a higher ROC (i.e., higher discriminability) than the presentation of faces in isolation.

We do not mean to suggest that the available evidence conclusively supports either the signal-detection model of eyewitness identification or the diagnostic feature-detection hypothesis. Conceivably, a threshold model of recognition memory will eventually be found to outperform a signal-detection model, and a principle other than the diagnostic feature-detection hypothesis will eventually be found to explain why simultaneous lineups yield higher discriminability than showups and sequential lineups. We advance these ideas – both of which are grounded in longstanding principles drawn from the experimental psychology literature (e.g., Egan, 1958; Gibson, 1969) – not as settled knowledge but to fill what we perceive to be a theoretical vacuum in the domain of eyewitness identification and to motivate research on the key issue of discriminability.

References

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

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

Clark, S. E. & Godfrey, R. D. (2009). Eyewitness identification evidence and innocence risk. *Psychonomic Bulletin & Review, 16,* 22-42.

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

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

Ebbesen, E. B., & Flowe, H. D. (2002). Simultaneous v. sequential lineups: What do we really know? http://www2.le.ac.uk/departments/psychology/ppl/hf49/SimSeq%20Submit.pdf.

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

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

Godfrey, R. G. & Clark, S. E. (2010). Repeated Eyewitness Identification Procedures: Memory, Decision Making, and Probative Value. *Law and Human Behavior, 34*, 241-258.

Goodsell, Gronlund & Carlson (2010). Exploring the Sequential Lineup Advantage Using WITNESS. *Law and Human Behavior, 34*, 445-459.

Gronlund, S. D., Carlson, C. A., Dailey, S. B., & Goodsell, C. A. (2009). Robustness of the sequential lineup advantage. *Journal of Experimental Psychology: Applied, 15,* 140-152.

Gronlund, S. D., Carlson, C. A., Neuschatz, J. S., Goodsell, C. A., Wetmore, S., 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. (in press). Evaluating Eyewitness Identification Procedures Using ROC Analysis. *Current Directions in Psychology Science.*

Innocence Project (2013). Understand the causes: the causes of wrongful conviction. New York: Innocence Project. http://www.innocenceproject.org/understand. Accessed June 7, 2013.

Klauer, K. C., & Kellen, D. (2012). The Law of Categorical Judgment (Corrected) extended: A note on Rosner and Kochanski (2009). *Psychological Review, 119,* 216-220.

Lindsay, R. C. L. (1999). Applying Applied Research: Selling the Sequential Line-up. Applied *Cognitive Psychology, 13,* 219-225.

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

Macmillan, N. A., & Kaplan, H. L. (1985). Detection theory analysis of group data: Estimating sensitivity from average hit and false-alarm rates. *Psychological Bulletin, 98,* 185-199.

Macmillan N. A. & Creelman, C. D. (1996). Triangles in ROC space: History and theory of "nonparametric" measures of sensitivity and response bias. *Psychonomic Bulletin & Review, 3,* 164-170.

Macmillan N. A. & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). Mahwah, NJ: Erlbaum.

Metz, C. E. (2006). Receiver operating characteristic analysis: a tool for the quantitative evaluation of observer performance and imaging systems. *Journal of the American College of Radiology, 3,* 413-422.

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

Mundy, M. E., Honey, R. C., & Dwyer, D. M. (2007). Simultaneous presentation of similar stimuli produces perceptual learning in human picture processing. *Journal of Experimental Psychology: Animal Behavior Processes, 33,* 124-138.

Myles-Worsley, M., Johnston, W. A., & Simons, M. A. (1988). The influence of expertise on X-ray image processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 553–557.

Ogilvie, J. C. & Creelman, C. D. (1968). Maximum-likelihood estimation of receiver operating characteristic parameters. *Journal of Mathematical Psychology, 5,* 377-391.

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

Steblay, N. K., Dysart, J., Fulero, S., & Lindsay, R. C. L. (2001). Eyewitness accuracy rates in sequential and simultaneous lineup presentations: A meta-analytic comparison. *Law and Human Behavior, 25,* 459-473.

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.

Swets J. A. (1996). *Signal detection theory and ROC analysis in psychology and diagnostics: collected papers*. Mahwah, NJ; Lawrence Erlbaum Associates.

Technical Working Group for Eyewitness Evidence. (1999). *Eyewitness evidence: A guide for law enforcement*. Washington, DC: U.S. Department of Justice, Office of Justice Programs.

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

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

Wells, G. L., Lindsay, R. C. L., & Ferguson, T. J. (1979). Accuracy, confidence, and juror perceptions in eyewitness identification. *Journal of Applied Psychology, 64*, 440–448.

Wells, G. L. & Quinlivan, D. S. (2009). The Eyewitness Post-Identification Feedback Effect: What is the Function of Flexible Confidence Estimates for Autobiographical Events? *Applied Cognitive Psychology, 23*, 1153-1163.

Wells, G. L., Small, M., Penrod, S., Malpass, R. S., Fulero, S. M. & Brimacombe, C. A. E. (1998). Repeated Eyewitness Identification Procedures: Memory, Decision Making, and Probative Value. *Law and Human Behavior, 22*, 1-39.

Wixted, J. T. & Mickes, L. (2012). The field of eyewitness memory should abandon "probative value" and embrace Receiver Operating Characteristic analysis. *Perspectives on Psychological Science, 7*, 275-278.

Zarkadi, T., Wade, K. A., & Stewart, N. (2009). Creating fair lineups for suspects with distinctive features. *Psychological Science, 20,* 1448-1453.

Footnotes

[1]Any detailed signal-detection model of group data would need to contend with complicated issues that can arise from the fact that different eyewitnesses place their confidence criteria at different locations on the memory strength axis – issues that have been considered when signal-detection theory has been used in other contexts (e.g., Klauer & Kellen, 2012). However, despite complexities like these, signal-detection theory typically has high heuristic value when applied to group data.

[2]Though, unfortunately, courts of law often take into account high confidence IDs made in the courtroom (after those ratings may have become inflated) rather than high confidence IDs made at the time of the initial identification (which is the confidence rating that we are referring to).

[3]For the sake of simplicity, we make the assumption that the features in question were not verbalized at the time of encoding or, if they were, were not influenced by any verbal overshadowing effect.

Table 1

Memory strength values of four facial features (f1through f4) that are summed to yield an aggregate memory strength value for a face in a showup and in a simultaneous lineup.

Procedure	Suspect	Parameter	f1	f2	f3	f4	Σ	
Showup	Innocent	$\mu_{Innocent}$	1	1	0	0	2	d_a = 0.84
		$\sigma^2_{Innocent}$	1.56	1.56	1	1	5.12	
	Guilty	μ_{Guilty}	1	1	1	1	4	
		σ^2_{Guilty}	1.56	1.56	1.56	1.56	6.24	
Simultaneous Lineup	Innocent	$\mu_{Innocent}$			0	0	0	d_a = 1.25
		$\sigma^2_{Innocent}$			1	1	2	
	Guilty	μ_{Guilty}			1	1	2	
		σ^2_{Guilty}			1.56	1.56	3.12	

Note. f1 = race; f2 = age; f3 = shape of face; f4 = size of eyes; d_a = measure of discriminability

Figure Captions

Figure 1. A depiction of the standard Unequal-Variance Signal-Detection (UVSD) model for three different levels of response bias: liberal bias (**A**), neutral (**B**) and conservative bias (**C**). In all three panels, the mean and standard deviation of the lure distribution are 0 and 1, respectively, and the mean and standard deviation of the target distribution are both 1.5. The receiver operating characteristic to the right shows the 3 ROC points that correspond to the 3 signal detection models shown to the left. The ROC point labeled "**A**" corresponds to the model shown at the top left; the ROC point labeled "**B**" corresponds to the model shown at the middle left, and the ROC point labeled "**C**" corresponds to the model shown at the bottom left.

Figure 2. Signal-detection models associated with two hypothetical eyewitness memory procedures (Procedure A and Procedure B) and the corresponding receiver operating characteristics associated with those models. Procedure A facilitates the discrimination between innocent and guilty suspects (less distributional overlap and correspondingly higher ROC) to a greater extent than Procedure B (greater distributional overlap and correspondingly lower ROC). The diagonal line in the ROC plot represents chance performance.

Figure 3. Receiver Operating Characteristic data comparing simultaneous vs. sequential lineups. The data are from Mickes et al. (2012).

Figure 1

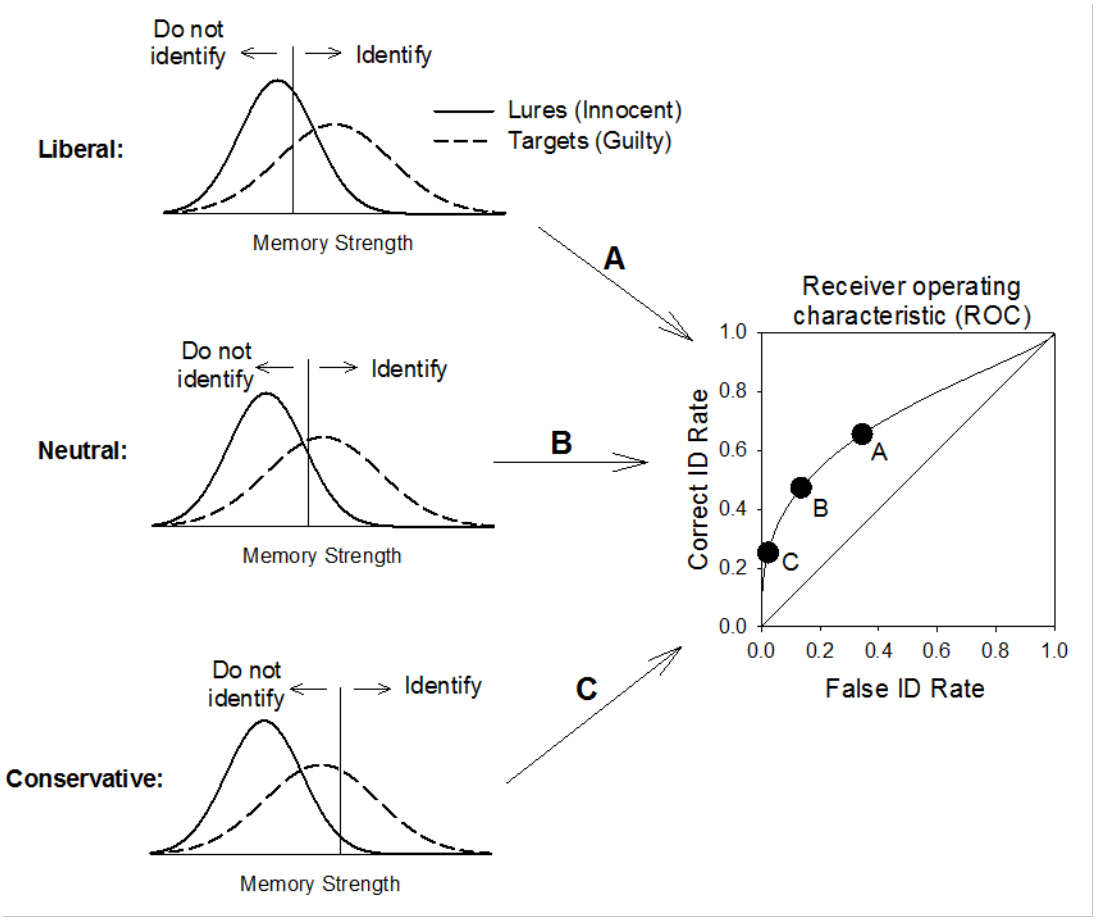

Liberal:
Do not identify
Identify
Lures (Innocent)
Targets (Guilty)
Memory Strength
A
Neutral:
Do not identify
Identify
Memory Strength
B
Conservative:
Do not identify
Identify
Memory Strength
C
Receiver operating characteristic (ROC)
Correct ID Rate
False ID Rate
0.0
0.2
0.4
0.6
0.8
1.0
A
B
C

Figure 2

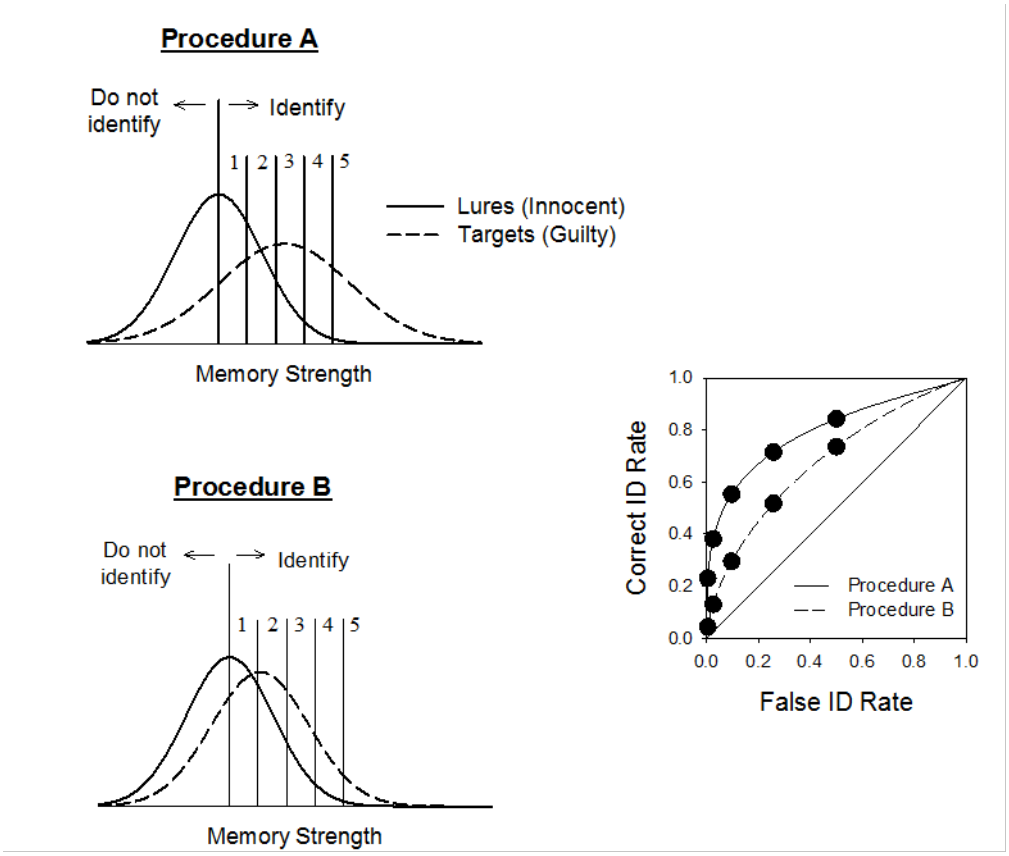
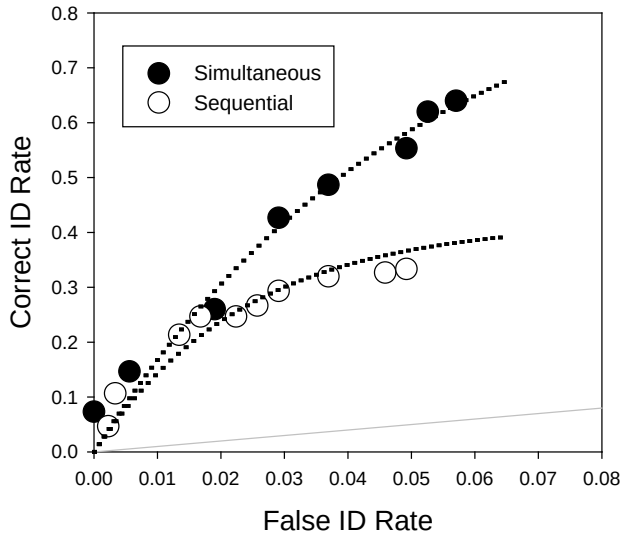

Procedure A
Do not identify
Identify
1 2 3 4 5
Lures (Innocent)
Targets (Guilty)
Memory Strength
Procedure B
Do not identify
Identify
1 2 3 4 5
Memory Strength
Correct ID Rate
False ID Rate
Procedure A
Procedure B
0.0 0.2 0.4 0.6 0.8 1.0

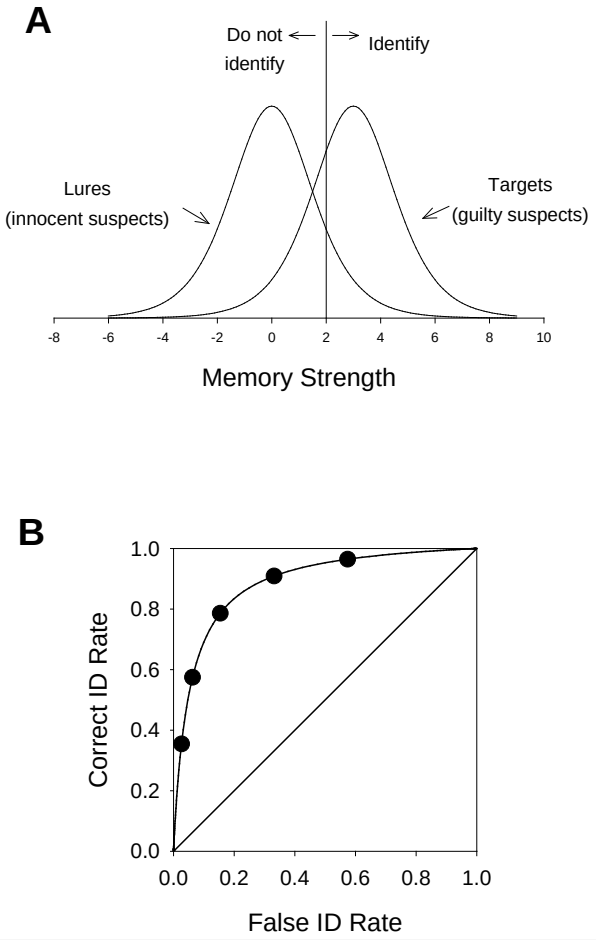
Figure 3


Simultaneous
Sequential
Correct ID Rate
False ID Rate
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08

Appendix

The logistic distribution can be used as an approximation to the Gaussian in order to easily illustrate that the diagnosticity ratio increases monotonically as the criterion sweeps from a liberal to a conservative setting (holding discriminability constant). We use the logistic for two reasons: (1) it is more convenient mathematically because the predictions can be illustrated using algebraic formulas (Ogilivie & Creelmen, 1968, used the logistic for this reason as well when they first explained how to fit a signal detection model to data using maximum likelihood estimation), and (2) to underscore the fact that our argument in favor of using signal-detection theory to guide thinking about eyewitness identification does not strongly depend on the assumption that the underlying memory strength distributions are specifically Gaussian in form.

Figure A1



For the sake of simplicity, we set the logistic parameter that governs the standard deviation to 1 (so that it drops out of the equation) and assume an equal-variance model. This logistic signal-detection model is shown in Figure A1A, and the criterion, *c*, placed at 2 on the memory strength axis (i.e., $c = 2$). Lower values of *c* correspond to a more liberal setting, with the most liberal setting being $-\infty$ (far to the left in Figure A1A). Similarly, higher values of *c*

correspond to a more conservative setting, with the most conservative setting being ∞ (far to the right in Figure A1A). Figure A1B shows the ROC that corresponds to the logistic signal-detection model shown in Figure A1A.

The probability density function for the logistic target distribution is:

$$p(x|target) = \frac{e^{-(x-\mu)}}{[1+e^{-(x-\mu)}]^2}$$

where x is a memory strength value associated with a particular target. The corresponding probability density function for the logistic lure distribution (obtained by setting $\mu = 0$) is:

$$p(x|lure) = \frac{e^{-x}}{[1+e^{-x}]^2}$$

The correct ID rate (*CID*) is given by the area under the target distribution to the right of the criterion, which is simply the integral of the logistic target distribution from *c* to ∞:

$$CID = \int_c^{\infty} \frac{e^{-(x-\mu)}}{[1+e^{-(x-\mu)}]^2}\, dx$$

which, as noted by Ogilvie and Creelman (1968), is equal to:

$$CID = \frac{1}{1+e^{c-\mu}} \quad (1)$$

Similarly, the false ID rate (*FID*) is given by the area under the lure distribution to the right of the criterion, which is the integral of the logistic lure distribution from c to ∞:

$$FID = \int_c^{\infty} \frac{e^{-x}}{[1+e^{-x}]^2}\, dx$$

which is equal to:

$$FID = \frac{1}{1+e^{c}} \quad (2)$$

Our goal is to determine what this logistic signal-detection model predicts about the diagnosticity ratio as the setting of *c* becomes more conservative. The diagnosticity ratio is defined as:

$$Diagnosticity\ Ratio = \frac{CID}{FID}$$

According to Equations 1 and 2 above, the diagnosticity ratio is given by:

$$\frac{CID}{FID} = \frac{1 + e^{c}}{1 + e^{c-\mu}}$$

or, equivalently:

$$\frac{CID}{FID} = \frac{1 + e^{c}}{1 + e^{-\mu}e^{c}} \tag{3}$$

In Equation 3, it is easy to see that the diagnosticity ratio increases as the criterion becomes more conservative. For example, at the most liberal setting, c = -∞ (i.e., *c* is set as far to the left as possible). In that case, $e^{c} = 0$ and Equation 3 reduces to 1:

$$\frac{CID}{FID} = \frac{1 + 0}{1 + e^{-\mu} \cdot 0} = \frac{1}{1} = 1$$

That is, when responding is as liberal as possible, the diagnosticity ratio is as low as possible. A more conservative but still liberal setting of *c* would be to place it at 0 (i.e., at the mean of the lure distribution). In that case, $e^{c} = 1$ and Equation 3 reduces to:

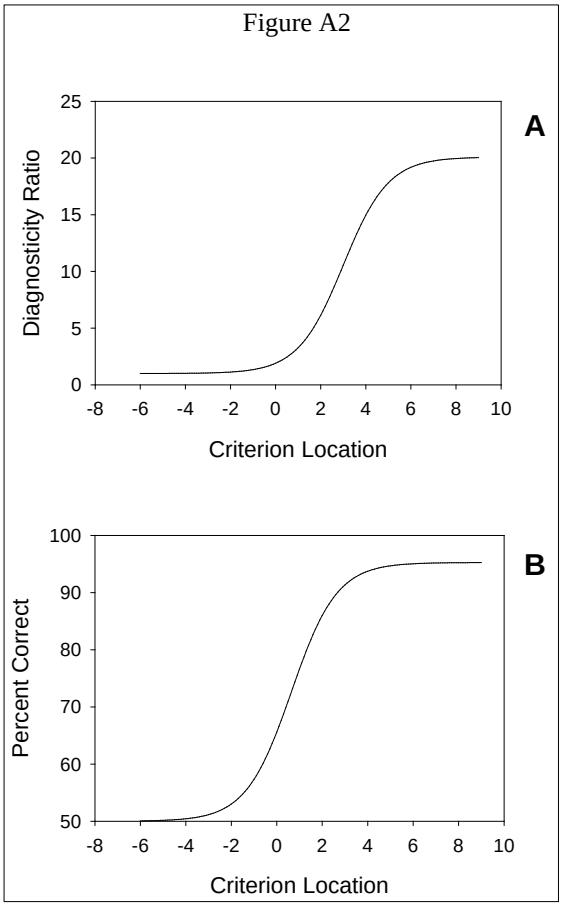
$$\frac{CID}{FID} = \frac{1 + 1}{1 + e^{-\mu} \cdot 1} = \frac{2}{1 + e^{-\mu}}$$

In the example shown in Figure A1A, $\mu = 3$, so $e^{-\mu} = 0.05$. Thus, the diagnosticity ratio at this more conservative setting of the criterion increases to 2 / (1 + 0.05) ≈ 2 / 1 = 2. At the most conservative setting, c = +∞ (i.e., the criterion is set as far to the right as possible). In that case, $e^{c} = \infty$ and Equation 3 reduces to:

$$\frac{CID}{FID} = \frac{1}{e^{-\mu}} = e^{\mu}$$

Because $\mu = 3$ in this example, $e^{\mu} = 20.1$, which is to say that the diagnosticity ratio at the most conservative setting is 20.1. Thus, as the criterion sweeps from -∞ to +∞ (i.e., as it sweeps from the most liberal setting to the most conservative setting), the diagnosticity ratio increases from 1 to 20.1. Figure A2A shows the continuous relationship between the diagnosticity ratio (CID / FID, which is plotted on the y-axis) and *c* (which is plotted on the x-axis) as *c* ranges from the liberal value of -6 to the very conservative value of +9. In other words, Figure A2A is a plot showing how Equation 3 behaves when $\mu = 3$. Clearly, the diagnosticity ratio spans a very wide range even though the ability to discriminate targets from foils (represented by the overlap of the two distributions shown in Figure A1A) remains constant. If an equal-variance Gaussian model were used to generate predictions instead, the diagnosticity ratio would continue to increase to infinity as the criterion moved ever further in the conservative direction. The key point is that both distributions (and many other continuous distributions as well) predict that the diagnosticity ratio increases as responding becomes more conservative. Figure A2B plots percent correct on the y-axis, where percent correct = 100% * *CID* / (*CID* + *FID*). Obviously, the model predicts that accuracy ranges from chance (50% correct) to near perfect as the criterion sweeps from left to right (liberal to conservative). Thus, the signal-detection model predicts that

Figure A2



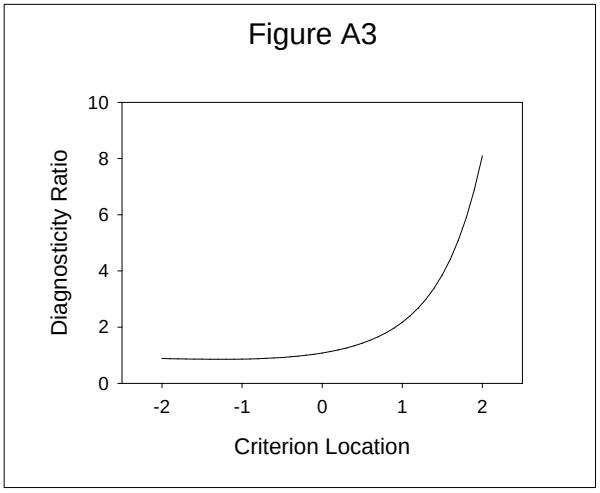
confidence and accuracy should be strongly related (because different levels of confidence reflect different locations of the criterion ranging from liberal to conservative).

Here, it is useful to differentiate two ways in which the word accuracy is used in the diagnostic context. In the sense just described, accuracy (percentage correct) increases as confidence increases or, equivalently, as responding becomes more conservative. This is true of both simultaneous and sequential lineups. Thus, an eyewitness memory procedure that induces more conservative responding (such as the sequential lineup procedure) might yield more accurate performance than a procedure that induces more liberal responding (such as the simultaneous lineup) when the focus is placed solely on overall correct and false ID rates (which include IDs made with any level of confidence). However, while true, this would not mean that the procedure that yields more conservative responding is the diagnostically more accurate procedure. If the correct and false ID rates for the two procedures fall on the same ROC, then they are equally accurate diagnostic procedures. The reason is that the same performance levels could be achieved using either procedure simply by adjusting the decision criterion. The diagnostically more accurate procedure is the one that yields a higher correct ID rate when the false ID rates for the two procedures are equated. In other words, the diagnostically more accurate procedure is the one that yields the higher ROC.

Equation 3 also makes it easy to see that the diagnosticity ratio increases with discriminability as well. In the equal-variance logistic model, discriminability increases as μ increases (i.e., distributional overlap decreases as μ increases). When $\mu = 0$, Equation 1 reduces to 1. When $\mu = \infty$, Equation 1 reduces to $1 + e^c$, which is obviously a number greater than 1. Because the diagnosticity ratio is sensitive to both discriminability and response bias, it is not a particularly informative measure for capturing either signal-detection property.

The predictions offered above were based on an equal variance logistic model, but the standard model for recognition memory is an *unequal* variance signal detection model that typically assumes Gaussian distributions. Does the Gaussian UVSD model also predict that the diagnosticity ratio will increase monotonically as responding becomes more conservative? For all practical purposes, it does. However, when the variance of the target distribution exceeds the variance of the lure distribution, the relationship can become non-monotonic at extreme liberal settings of the criterion. The non-monotonicity becomes more apparent the more the distributions differ in variance and the lower the level of discriminability is. As a general rule, the variances of the target and lure distributions tend to become equal as discriminability approaches zero, so the issue is not likely to be relevant in practice. However, even if the variances remained very unequal when discriminability was low, the non-monotonicity would occur at such an extremely liberal setting of the criterion that the issue would still probably not be relevant in practice. Consider, for example, a Gaussian signal detection model with means that are separated by only 0.2 standard deviation units, with the lure distribution having a standard deviation of 1 and the target distribution having a standard deviation of 2 (a standard deviation difference that would be unusually large even if discriminability were high). Figure A3 shows the predicted diagnosticity ratio as the criterion sweeps from 2 standard deviations below the mean of the lure distribution (a setting so liberal that an ID would almost always be made) to 2 standard deviations above the mean of the target distribution. The predicted diagnosticity ratio at -2 is 0.88, which is slightly greater than the predicted minimum

Figure A3



diagnosticity ratio of 0.86, which occurs at -1.3 (i.e., that is the inflection point of the non-monotonic relationship). From that point on, the diagnosticity ratio increases monotonically, and it would continue to increase towards infinity as the criterion were placed at increasingly conservative points beyond the upper limit shown in Figure A3. Thus, even for an extreme model like this, Figure A3 shows that the UVSD model predicts that the diagnosticity ratio will increase over a very wide range of reasonable settings as responding becomes more conservative. The relationship only becomes more monotonic as more plausible values for the UVSD model are used (i.e., the slight non-monotonicity evident in Figure A3 becomes increasingly negligible as more plausible UVSD parameters are used). Thus, for all practical purposes, the standard UVSD model predicts that the diagnosticity ratio will increase as responding becomes more conservative.