Models of lineup memory

Article *in* Cognitive Psychology · July 2018

DOI: 10.1016/j.cogpsych.2018.06.001

John Wixted
University of California, San Diego
164 PUBLICATIONS **9,870** CITATIONS

Laura Mickes
Royal Holloway, University of London
65 PUBLICATIONS **1,600** CITATIONS

Models of Lineup Memory

John T. Wixted[1], Edward Vul[1], Laura Mickes[2] & Brent M. Wilson[1]

[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. Ed Vul, Department of Psychology, University of California, San Diego. Laura Mickes, Department of Psychology, Royal Holloway, University of London. Brent Wilson, Department of Psychology, University of California, San Diego.

This work was supported by the National Science Foundation SES-1456571 to John T. Wixted. 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).

Abstract

Face recognition memory is often tested by the police using a photo lineup, which consists of one suspect, who is either innocent or guilty, and five or more physically similar fillers, all of whom are known to be innocent. For many years, lineups were investigated in lab studies without guidance from standard models of recognition memory. More recently, signal detection theory has been used to conceptualize lineup memory and to motivate receiver operating characteristic (ROC) analysis of competing lineup procedures. However, this movement is still in its infancy. Here, we present three competing signal-detection models of lineup memory, derive their likelihood functions, and fit them to empirical ROC data. We also introduce the notion that the memory signals generated by the faces in a lineup are likely to be correlated because, by design, they share many features. The models we investigate differ in their predictions about the effect that correlated memory signals should have on the ability to discriminate innocent from guilty suspects. The best-fitting model incorporates a principle known as "ensemble coding," a concept that applies to the presentation of any set of similar items (including the faces in a lineup). The ensemble model also accords with a previously proposed theory of eyewitness identification according to which the simultaneous presentation of faces in a lineup enhances discriminability compared to when faces are presented in isolation because it permits eyewitnesses to detect and discount non-diagnostic facial features.

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

Models of Lineup Memory

Eyewitness misidentifications have contributed to a large number of wrongful convictions, and laboratory-based research designed to reduce that problem has focused largely on the format of lineups that the police use during the early stages of a criminal investigation (e.g., Lindsay & Wells, 1985). For many years, the relevant data were analyzed without any reference to the conceptual and analytical tools that are commonly used by cognitive psychologists to study recognition memory, but more recent research differs in that it has relied on signal detection theory to conceptualize and analyze receiver operating characteristic (ROC) data. However, thus far, competing signal detection models of eyewitness identification have not been formally specified and then tested for their ability to accurately characterize empirical data. The purpose of this article is to do just that.

Although live lineups were once the norm, nowadays ~90% of lineups administered by the police in the U.S. are photo lineups (Police Executive Research Forum, 2013). Like a live lineup, a photo lineup consists of one suspect, who is either innocent or guilty, and several (usually five) physically similar fillers, all of whom are known to be innocent. Typically, the photos are presented simultaneously to the witness, who can (1) identify the suspect (suspect ID), (2) identify a filler (filler ID), or (3) reject the lineup (no ID). Alternatively, the photos can be presented sequentially, with the procedure terminating when the first positive ID is made (Lindsay & Wells, 1985). Here, we focus on theories of recognition memory tested using the simultaneous photo-lineup procedure.

The lineup task is similar to a list-memory recognition task in many ways, but an important difference is that in a list-memory design, many different items are tested with one participant. By contrast, in a common lineup design, many different participants are tested with one set of items. Thus, instead of different items contributing to the variance of the distribution of memory signals across trials, different participants do. Either way, one distribution of memory signals is generated by previously seen targets, and the other is generated by novel lures. In an eyewitness identification experiment, the targets are guilty suspects, and the lures are innocent suspects and fillers. Achieving a greater theoretical understanding of those two memory-strength distributions is the goal of this article, and we do so by testing the ability of three specific signal-detection-based models of lineup memory to quantitatively characterize empirical ROC data.

Background Theoretical Considerations

Before delving into modeling details, we consider several preliminary theoretical and empirical issues. First, we describe how lineup memory is generally conceptualized within a signal detection framework and how each of the three signal detection models we later consider is defined by its unique diagnostic memory-strength variable. We then briefly survey prior research on the diagnostic variable that participants appear to rely upon when memory is tested using a collection of test items. Lastly in this section, we introduce the key notion of correlated memory signals in lineups, the predicted effect of which differs depending on which model is correct.

Modeling Lineup Memory using Signal Detection Theory

The simplest signal detection model for simultaneous lineups was briefly mentioned by Macmillan & Creelman (1991, p. 251) in their classic signal-detection text and was considered in more detail by Duncan (2006) in a technical report. They both referred to this model as the Independent Observations model, as we will. According to this simple model, which we illustrate here in Figure 1, memory strength values for lures (innocent suspects and fillers) and for targets (guilty suspects) are distributed according to Gaussian distributions with means of μ_{Lure} and

μ_{Target}, and standard deviations of σ_{Lure} and σ_{Target}, respectively. The innocent suspect is, from the witness's point of view, just another filler (assuming, as we do, a fair lineup). Hence, there is only one lure distribution. A 6-member target-present lineup is conceptualized as 5 random draws from the lure distribution and 1 random draw from the target distribution, and a fair 6-member target-absent lineup is conceptualized as 6 random draws from the lure distribution. For the equal-variance case, $\sigma_{Target} = \sigma_{Lure} = \sigma$, the ability to discriminate innocent from guilty suspects (d'_{IG}) is given by $d'_{IG} = (\mu_{Target} - \mu_{Lure}) / \sigma$.

Note that, in the eyewitness context, d'_{IG} is a population measure of discriminability, not a measure of discriminability for any particular participant. Eyewitness identification studies in the laboratory often involve a large number of once-tested participants (e.g., $N = 1000$). Each participant, if tested individually using a list-memory procedure, would presumably yield a different d' score, reflecting the fact that some participants have better memories than others. The range of memory ability across once-tested participants in an eyewitness identification experiment is one of many possible sources of the variance represented by σ^2_{Target} and σ^2_{Lure}. Other possible sources of variance include (for example) how much attention participants paid to the mock-crime video and how similar the perpetrator in the video is to someone previously known to the participants.

In a signal detection model (Figure 1), confidence ratings correspond to different decision criteria. Assuming 5 different levels of confidence associated with an ID, there are 5 different confidence criteria. The parameters c_1 through c_5 in Figure 1 represent the confidence criteria for positive IDs of a suspect or a filler. According to the Independent Observations model, a witness first determines which face generates the strongest memory signal (the MAX face) and then identifies that face if its memory signal exceeds c_1, without regard for the memory signals generated by the other faces in the lineup. If the strength of the memory signal exceeds a higher criterion (e.g., c_3), the ID is made with correspondingly higher confidence. Although confidence ratings are sometimes taken when the decision is to reject the lineup, our focus is on predicting confidence ratings associated with positive IDs, which are made in relation to a particular lineup member and which are used to evaluate the reliability of eyewitness identifications in courts of law.

The model illustrated in Figure 1 may be the simplest signal detection model for simultaneous lineups, but it is by no means the only one. The three models we investigate in this article differ in their assumptions about how the memory signals generated by the faces in a lineup are used to decide whether or not to identify the face that generates the strongest signal. We refer to these three models as the Independent-Observations model (Figure 1), the Integration model, and the Ensemble model. The diagnostic memory-strength variable for the Independent-Observations model is the raw (untransformed) memory-match signal generated by a face in the lineup (Macmillan & Creelman, 2005); for the Integration model, it is the *sum* of the memory-match signals generated by the faces in the lineup (Duncan, 2006); and for the Ensemble model, it is the *difference* between the memory-match signal for a given face and the average of the memory signals generated by all of the faces in the lineup. As described below, this model is a mathematical instantiation of the diagnostic feature-detection theory proposed by Wixted and Mickes (2014).

All three models rely on a MAX decision rule. According to this rule, the face in the lineup that generates the strongest memory-match signal is identified if the relevant decision variable exceeds a criterion; otherwise, the lineup is rejected (i.e., no ID is made).[1] Table 1 summarizes, for each model, the decision variable (i.e., the variable that is considered in relation to the various confidence criteria) and the decision rule associated with the MAX face in the lineup. Note that for all three models, the face that generates the strongest memory signal is the only face that is a candidate for being identified. The larger the magnitude of the decision variable, the more likely it is that the MAX face will be identified and the higher the eyewitness's confidence in that ID will be. We focus on these three specific models because, as we show later, they have all been previously proposed (i.e., they are the currently competing signal detection models of lineup memory).

Prior Research on the Nature of the Decision Variable

What does prior research suggest about the nature of the diagnostic variable when memory is tested using a lineup? Studies from several domains that are relevant to this question have investigated the effect of adding implausible (i.e., "dud") alternatives to a set of test items on confidence in decisions about the plausible (i.e., non-dud) alternatives in the set. According to the Independent Observations model, confidence is theoretically determined by the memory signal associated with the MAX face without regard for the other faces in the lineup. Thus, all else being equal, the addition of duds (which are very unlikely to generate the MAX signal) should have no effect on confidence. Contrary to that prediction, in the context of multiple-choice general knowledge questions, Windschitl and Chambers (2004) found that the addition of implausible alternatives increased confidence in the plausible alternatives. Essentially the same result was found in an associative recognition task conducted by Hanczakowski, Zawadzka &

[1] The MAX rule is often assumed to apply in visual search tasks involving target-present and target-absent arrays (e.g., Cameron, Tai, Eckstein & Carrasco, 2004; Palmer, Fencsik, Flusberg, Horowitz & Wolfe, 2011; Palmer, Verghese & Pavel, 2000; Smith & Sewell, 2013; Verghese, 2001).

Higham (2014) and in eyewitness identification studies conducted by Charman et al. (2011) and Horry and Brewer (2016).

Windschitl and Chambers (2004) accounted for their finding in terms of a *contrast hypothesis* according to which adding duds increases the number of pairwise comparisons that strongly favor the most plausible alternative, thereby increasing confidence in that alternative. This account emphasizes the *difference* between the MAX (i.e., most plausible) item and the other items in the set, which is most similar to the diagnostic variable envisioned by the Ensemble model. Hanczakowski et al. (2014) extended this account based on Tversky's (1977) idea that the local context determines which features in a set of stimuli are considered diagnostic for the task at hand. This account holds that it is the difference between a contextually plausible item — which is differentially associated with diagnostic features — vs. the other items in the set that serves as the diagnostic decision variable.

Wixted and Mickes (2014) applied the same idea to lineup memory to explain why simultaneous lineups often yield higher discriminability than sequential lineups. Their diagnostic feature-detection theory holds that the simultaneous presentation of faces helps witnesses to notice and to then discount non-diagnostic facial features (namely, the features that are common across the lineup members). By focusing on potentially diagnostic features, the memory signal of the guilty suspect (the lineup member whose diagnostic features most closely match the witness's memory of the perpetrator) will stand out from the crowd of signals associated with the other faces in the lineup. For present purposes, the key point is that these theoretical accounts all focus on a difference variable, just as the Ensemble model does.

In light of these considerations, it seems fair to suggest that the prior odds favor the Ensemble model over the competing signal detection models. Then again, as noted by

Hanczakowski et al. (2014), the effects of duds on confidence could be explained without assuming a difference variable by instead assuming a criterion shift. According to this idea, when duds are added to the set, a more liberal decision criterion is used to express high confidence, a possibility they termed “recalibration.” One way to distinguish between that possibility and the alternative possibility that the diagnostic memory-strength variable truly consists of a difference variable is to fit the three competing models to empirical ROC data, which is what we do here.

Correlated Memory Signals

A final background theoretical issue to consider is the role of correlated memory signals in lineups. The memory signals in a lineup are likely to be correlated by virtue of the fact that the standard approach to creating lineups is to select one suspect and 5 fillers that correspond to the physical description of the perpetrator provided by the eyewitness. If the lineup-defining features of the perpetrator happen to generate a strong memory signal, then, because those features will be shared by everyone in the lineup, all of the faces in the lineup – not just the face of the perpetrator – will tend to generate a relatively strong memory signal as well (i.e., the mean memory signal of the faces in the lineup would be high). This might happen, for example, if the witness described the perpetrator as having a flamboyant handlebar mustache and shocking red hair, in which case everyone in the lineup would have those memorable features. By contrast, if the lineup-defining features associated with the perpetrator are not particularly distinctive (or if the features are distinctive but the witness did not get a good look at the perpetrator), the relevant features would likely be weakly encoded and would therefore generate a weak memory signal. In that case, all of the faces in the lineup would tend to generate a weak memory signal as well (i.e., the mean memory signal of the faces in the lineup would all be low).

Although the effect of correlated memory signals on lineup memory has not been previously considered, the fact that correlated signals can facilitate performance in 2-alternative forced-choice (2AFC) recognition memory has long been known (Hall, 1979; Hintzman, 1988, 2001; Tulving, 1981). In a 2AFC task, participants are presented with a forced choice between a previously studied target and a novel lure, and they are instructed to choose the item that they believe to be the target. The optimal strategy on the 2AFC task is to base the decision on the difference between the memory signals generated by the target and the lure (Macmillan, 2002). Thus, for example, on every trial, the participants might compute the memory-strength difference between the item on the right and the item on the left and then choose the right item if the result is positive and choose the left item if the result is negative. For the typical case in which the memory strength of a target does not predict the memory strength of the lure (i.e., for the typical case in which the memory signals of the targets and lures are uncorrelated), the subtraction process would give rise to two distributions, one with a mean and variance of μ_{Target} and σ^2_{Target} + σ^2_{lure}, respectively, and the other with a mean and variance of -μ_{Target} and σ^2_{Target} + σ^2_{lure}, respectively. Discriminability on the 2AFC task is given by the difference between the two means divided by their common standard deviation, or $d' = 2\mu_{Target}/\sqrt{(\sigma^2_{Target} + \sigma^2_{Lure})}$. The variances of the targets and lures in the denominator sum because when adding or subtracting uncorrelated random variables, the variance of the resulting random variable is the sum of the component variances.

On some 2AFC tasks, the targets are paired with similar lures, in which case the memory signals of the targets and lures would be correlated. For example, a target might be a picture of a previously presented violin, and its corresponding similar lure might be a picture of a novel violin that shares many features with the target. Because the two test items share many features, the memory-strength signals they generate will be correlated. Under those conditions, $d' = 2\left(\mu_{Target} - \mu_{Lure}\right)/\sqrt{\left(\sigma^2_{Target} + \sigma^2_{Lure} - 2\rho\sigma_{Target}\sigma_{Lure}\right)}$, where ρ is the correlation coefficient (Hintzman, 1988, 2001). We can simplify this equation by setting μ_{Lure} = 0 and by assuming an equal-variance model such that σ^2_{Target} = σ^2_{lure} = 1. In that case, discriminability on the correlated 2AFC task becomes $d' = \sqrt{2}\,\mu_{Target}/(1-\rho)$. This equation makes it clear that discriminability increases as ρ increases. As ρ approaches a perfect correlation of 1, discriminability tends to infinity. A perfect correlation means that the memory strength signals generated by the target and the lure on any given trial fall at precisely the same point on their respective distributions (Hintzman, 2001). For example, if the target on a given trial happens to fall 1 standard deviation below μ_{Target}, then the lure will fall one standard deviation below μ_{Lure}. So long as μ_{Target} > μ_{Lure}, when ρ = 1, the participant will correctly choose the target every time.

We propose that correlated memory signals play a potentially important role not only in in 2AFC recognition memory but also in lineup memory. Moreover, there are two distinct parts to the story of how correlated memory signals may affect lineup memory. The first part is independent of the three models we consider, whereas the second part is model specific (i.e., the effect is different for each model). The first part of the story concerns the beneficial effect of correlated memory signals on target-present lineup performance, which happens to be the same benefit that occurs for the 2AFC task. For example, in a target-present lineup, if ρ = 1, so long as μ_{Target} > μ_{Lure}, the memory-strength signal generated by the target (the guilty suspect) will exceed the memory-strength signals generated by the lures every time. In other words, the target will always be the MAX face in the lineup. Because all three models assume a MAX decision rule, when ρ = 1, only the target would be a candidate for identification in target-present lineups. If its memory strength exceeds the decision criterion, the target will be correctly identified, but no lure would ever be incorrectly identified. Thus, as described in more detail later, for all three models, d'_{TP} (i.e., the ability to discriminate the guilty suspect from the lures in a target-present lineup) tends to infinity as ρ approaches 1. By contrast, in fair target-absent lineups, the ability to discriminate the innocent suspect from the fillers is, by definition, equal to 0 (i.e., $d'_{TA} = 0$) because the innocent suspect is effectively another filler.

Lineup performance is not determined solely by what happens on target-present trials but also by what happens on target-absent trials. Thus, to fully predict lineup performance across all lineups, it is also important to consider d'_{IG}, the aggregate ability of eyewitnesses to discriminate innocent suspects in target-absent lineups from guilty suspects in target-present lineups. Unlike d'_{TP}, which always increases as ρ increases, the effect of correlated memory signals on d'_{IG} differs for the three competing models under consideration here. This is the second part of the story of correlated memory signals on lineup memory. As described next, a positive correlation between memory signals in a lineup should have no further effect on d'_{IG} according to the Independent Observations model, it should exert a *negative* effect according to the Integration model (decreasing d'_{IG}), and it should exert a further *positive* effect (increasing d'_{IG}) according to the Ensemble model. Critically, it is the combined effect on d'_{TP} and d'_{IG} (two measures of underlying discriminability) that determines the effect that correlated memory signals have on the empirical discriminability as revealed by ROC data (Wixted & Mickes, 2018).

Three Models of Lineup Memory

In this section, we formally derive the predictions that these three models make about d'_{IG} (the ability to discriminate innocent from guilty suspects) when memory signals are uncorrelated and when they are correlated. After the models are formally specified, in subsequent sections, we derive the likelihood functions for each model and then fit the models to empirical ROC data.

Independent Observations Model

According to the Independent Observations model, the memory signals generated by innocent and guilty suspects (and fillers) are considered without regard for the memory signals generated by the other faces in the lineup. This simple model is similar to the BEST model implemented by Clark (2003; Clark, Erickson & Breneman, 2011) in the context of the WITNESS model and has often been used to frame a recent debate about the utility of ROC analysis in eyewitness identification (e.g., Lampinen, 2016; Rotello & Chen, 2016; Smith, Wells, Lindsay, & Penrod, 2017; Wixted, Mickes, Wetmore, Gronlund & Neuschatz, 2017). As depicted in Figure 1, the mean signal generated by guilty suspects and innocent suspects would be μ_{Target} and μ_{Lure}, respectively, and their corresponding standard deviations would be σ_{Target} and σ_{Lure}, respectively. Note that here and throughout this article, μ_{Target} and μ_{Lure} represent the means of the raw (untransformed) memory distributions for targets and lures, respectively. Assuming an equal-variance model ($\sigma_{Target} = \sigma_{Lure} = \sigma$), discriminability based on performance aggregated across all lineups is given by $d'_{IG} = (\mu_{Target} - \mu_{Lure}) / \sigma$.[2] By convention, we set $\mu_{Lure} = 0$, so the numerator of the d'_{IG} equation reduces to μ_{Target}, and we set $\sigma = 1$ in the denominator, so the ability to discriminate innocent from guilty suspects reduces to the simple equation:

$$d'_{IG} = \mu_{Target} \tag{1}$$

The Independent Observations model further assumes that on a given trial, a decision is based on the face in the lineup that generates the maximum (MAX) memory strength signal.

[2] We assume an equal-variance model mainly for simplicity. List-memory studies of recognition memory usually support an unequal-variance model (greater variance for the target distribution), but, as we show later, lineup data are often consistent with an equal-variance model (or an unequal-variance model in the opposite direction, with greater variance for lures).

Our concern for the moment is what the Independent Observations model predicts about the memory-strength signals generated by targets and lures (i.e., about d'_{IG}) *before* a decision is made using the MAX rule. The specific question of interest is how d'_{IG} should be affected by the presence of positively correlated memory signals according to this model, and the answer is that it should not be affected at all. The reason is that, according to this model, the memory signals generated by the faces in the lineup are considered without regard for the memory signals generated by the other faces in the lineup. Regardless of the size of the correlation, the targets are drawn from a distribution with mean μ_{Target} and standard deviation σ_{Target}, and the foils are drawn from a distribution with mean μ_{Lure} and standard deviation σ_{Lure}. Thus, in the equal-variance case ($\sigma_{Target} = \sigma_{Lure} = \sigma$), $d'_{IG} = (\mu_{Target} - \mu_{Lure}) / \sigma$, and this is true whether ρ equals 0 or 1 or anything in between. Setting $\mu_{Lure} = 0$ and σ, = 1, this equation reduces to Equation 1 regardless of ρ.

Using simulated data, Figure 2 illustrates the effect of increasingly correlated memory signals for the Independent Observations model in which $\mu_{Target} = 2$, $\mu_{Lure} = 0$, and, assuming equal variance, $\sigma = 1$. The distributions in the top panel (Figure 2A) were generated by drawing values for innocent suspects/fillers from the lure distribution and guilty suspects from the target distribution with varying degrees of dependence. In the uncorrelated case, a value (y) drawn from the lure distribution, $y \sim N(\mu_{Lure}, \sigma)$, was independent of the value (x) drawn from the target distribution, $x \sim N(\mu_{Target}, \sigma)$. At the opposite extreme (correlation ≈ 1), the values of x and y for a given draw from their respective distributions were constrained such that $y - \mu_{Lure} = x - \mu_{Target}$. The resulting distributions in the top panel illustrate the fact that, for this model, innocent-vs.-guilty suspect discriminability (d'_{IG}) is unaffected by the size of the correlation. Again, keep in mind that these distributions represent the memory signals for innocent and guilty suspects across all lineups, whether or not they were the maximum values in the lineup on those trials.

Although the separation of these two distributions is the discriminability measure of interest, a suspect ID would have an opportunity to occur only on the subset of trials in which the suspect generated the MAX signal in the lineup.

Figure 2B shows the (sometimes skewed) distribution of memory signals on trials in which the innocent and guilty suspects in the above simulation were associated with the MAX signal in the lineup. In other words, these distributions show the subset of trials in which a suspect ID would occur if the strength of the MAX memory exceeded the decision criterion. Note that, as depicted here, these are not normalized distributions but are instead frequency distributions. They are plotted as frequency distributions to illustrate the increase in the absolute number of target-present trials in which the guilty suspect yields the MAX signal as the correlation increases (i.e., as d'_{TP} increases). This is evident in the fact that the height of the target-present MAX distribution – but not the target-absent MAX distribution – increases from left to right even though the number of simulated target-present and target-absent trials remains constant. The rightmost target distribution in Figure 2B is Gaussian with a mean of μ_{Target} and a standard deviation σ_{Target} because, on every target-present trial, the target generates the strongest memory signal.

When we later fit the Independent Observations model to empirical data, we will estimate d'_{IG} and the locations of the various confidence criteria in relation to the distributions aggregated across all lineups (not on the subset of lineups in which the suspect generates the MAX signal). That is, we quantify discriminability in terms of the distributions illustrated in Figure 1 (and in Figure 2A), not in terms of the extreme value distributions themselves. However, the likelihood functions we derive for the Independent Observations model make predictions about the probability of suspect IDs, filler IDs and no IDs based on the corresponding extreme value distributions, like the ones shown in Figure 2B. We derive the models and fit their parameters in relation to the distributions of memory signals aggregated across all lineups because the math is more tractable and the model fits are easier to interpret than would be the case if we based our analyses on the corresponding extreme-value distributions themselves.

Integration Model

The Integration model (Duncan, 2006; Macmillan & Creelman, 2005) assumes that the witness computes a *sum* of the memory signals generated by all of the faces in the lineup. If that summed value exceeds a decision criterion, then an ID will be made, otherwise the lineup is rejected. If the summed value exceeds a criterion, the specific face that is identified is the MAX face in the lineup. The Integration model has often been used in the eyewitness ID literature to conceptualize lineup memory or to compute *d'* for a lineup task (e.g., Duncan, 2006; Horry, Brewer, Weber & Palmer, 2015; Palmer & Brewer, 2012; Palmer, Brewer & Horry, 2013; Palmer, Brewer & Weber, 2010; Smith et al., 2017; in press). In fact, it seems fair to say that, at the present time, this is the dominant signal detection model in the field of eyewitness identification.

The Integration model is illustrated in Figure 3, which shows the distribution of summed memory-strength signals across all target-present and target-absent trials. The mean of the summed random variable on target-present trials (guilty suspect + fillers) is the sum of the means of the components, or μ_{Target} + $\sum_{1}^{k-1} \mu_{Lure}$ = μ_{Target} + (k-1)μ_{Lure}, where the sum reflects the fact that there are k – 1 fillers in the target-present lineup. On target-absent trials, the mean (innocent suspect + fillers) is simply μ_{Lure} + (k-1)μ_{Lure} = $k\mu_{Lure}$. Because we set μ_{Lure} = 0 by convention, the means of the summed memory-strength variables on target-present and target-absent trials are equal to μ_{Target} and 0, respectively. Thus, the difference between them (i.e., the numerator of the

d'_{IG} equation) is μ_{Target} - 0 = μ_{Target}, which is the same as the numerator of the d'_{IG} equation for the Independent Observations model. What differs is the denominator of the d'_{IG} equation because when random variables are summed, the variance of the constituent elements sum as well. For example, in a 2-person target-present lineup consisting of one target and one lure, the variance of the summed memory signal would be $\sigma^2_{Target} + \sigma^2_{Lure} + 2\rho\sigma_{Target}\,\sigma_{Lure}$. As shown in Appendix A, the variance for a summed variable in a lineup of size *k* (target-present or target-absent) is given by $k\sigma^2 + k(k\text{-}1)\rho\sigma^2$, which reduces to $k[1 + (k\text{-}1)\rho]$ after setting $\sigma^2 = 1$ and rearranging terms. Thus, the ability to discriminate innocent from guilty suspects according to the Integration model is given by:

$$d'_{IG} = \frac{\mu_{Target}}{\sqrt{k[1 + (k - 1)\rho]}} \tag{2}$$

This equation indicates that positively correlated memory signals should have the effect of *reducing* discriminability between innocent and guilty suspects compared to the uncorrelated case. The negative effect results from the fact that summing positively correlated random variables increases the variance of the summed variable beyond what it would otherwise be. Note that we have assumed that the computation of the summed variable is an error-free process. Predicted discriminability would be lower than is implied by Equation 2 if we added random error to the summation process, but the predictions of the model with respect to ρ would not otherwise be affected. That is, according to the Integration model, d'_{IG} decreases as ρ increases.

Figure 4A illustrates the distribution of the summed memory signals as envisioned by the Integration model as the correlation among memory signals increases from 0 to ~1 (once again based on a simulation in which, for the untransformed memory signals, $\mu_{Target} = 2$, $\mu_{Lure} = 0$, and $\sigma = 1$). For the untransformed signals, which are the signals used by the Independent Observations model, $d' = 2$ regardless of ρ (as illustrated earlier in Figure 2A). By contrast, for the transformed (summed) signals in Figure 4A, d'_{IG} is given by Equation 2, which means that, according to this model, the ability to discriminate innocent from guilty suspects decreases as the correlation increases. As a concrete example, for k = 6 and ρ = 0, $d'_{I\text{-}G} = 2/\sqrt{6[1+(6-1)0]} = 0.82$, but for ρ = .50, $d'_{I\text{-}G} = 2/\sqrt{6[1+(6-1).50]} = 0.44$. Thus, as is evident in Figure 4A, the overlap of summed memory signals associated with target-present from target-absent lineups increases (and d'_{IG} decreases) the more the memory signals are correlated.

The distributions in Figure 4B are frequency distributions of the summed decision variable on the subset of trials in which innocent or guilty suspects generated the MAX signal. Once again, these distributions show the increase in the absolute number of target-present trials in which the guilty suspect generates the MAX memory signal as the correlation increases (due to the increase in d'_{TP} for target-present lineups with increasing ρ). When we later fit the Integration model to empirical data, the parameters we estimate correspond to the parameters shown for the Integration model illustrated in Figure 3. Again, however, the likelihood functions make predictions about suspect IDs, filler IDs and no IDs from target-present and target-absent lineups based on the MAX distributions like the ones shown in Figure 4B.

Ensemble Model

The Ensemble model assumes that the subject computes the *difference* between the memory signal for each face and the average memory signal of all faces in the lineup. In essence, this decision variable corresponds to how much the memory signal for a given face stands out from the crowd of faces in the lineup. If the largest difference score exceeds a decision criterion, then the face associated with that difference score is identified (i.e., once again, a MAX rule is assumed).[3] This model is closely related to the BEST minus REST model implemented by Clark (2003; Clark et al., 2011) in the context of their WITNESS model.

The Ensemble model is grounded in an extensive body of recent research suggesting that when similar objects are presented together, summary statistics are quickly and automatically computed (Albrecht & Scholl, 2010; Ariely, 2001; Chong & Treisman, 2003). Such "ensemble coding" applies not only to a set of similar objects but also to a set of similar faces. For example, when shown pictures of four similar faces, subjects later recognize the mean identity (i.e., the morphed average of the presented faces) with a high probability (de Fockert & Wolfenstein, 2009; Neumann, Schweinberger, & Burton, 2013).

The Ensemble model is illustrated in Figure 5. Let x be a random variable for an individual face drawn from the target or lure distribution and y be a random variable drawn from the ensemble (average) distribution of a k-alternative lineup. The decision variable for the Ensemble model is $x – y$. On target-present trials, the mean of x for the target is equal to μ_{Target}, and the mean of y (the ensemble variable) is equal to ($\mu_{Target} + \sum_1^{k-1} \mu_{Lure}$) / k, where the sum corresponds to the $k – 1$ fillers in the lineup. Thus, the mean of the $x – y$ variable for the target is equal to μ_{Target} - ($\mu_{Target} + \sum_1^{k-1} \mu_{Lure}$) / k. Because $\mu_{Lure} = 0$, the mean of this difference score reduces to μ_{Target} - μ_{Target} / k = μ_{Target} (1 – 1/k).

The mean of the $x – y$ variable on target-absent trials is obtained by setting x for an individual filler equal to $\mu_{Lure} = 0$ (as we did for target-present fillers above) and by setting y equal to the mean of the 6 faces in the target-absent lineup. Because all $k – 1$ fillers and the

[3] Based on physical appearance alone, all of the faces in a fair lineup are plausible suspects for having committed the crime. However, if none of the faces were remotely plausible, the difference between the best face and the average face might still be large on a physical scale, but their psychological similarity to the perpetrator (i.e., the raw memory-match signal) would now be similarly small (Nosofsky, 1992).

innocent suspect in the lineup are drawn from a distribution with mean equal to μ_{Lure} = 0, the mean of the memory signals in the lineup equals 0. This result indicates that the mean of the $x - y$ variable for fillers on target-absent trials remains centered on 0.

The mean of the $x - y$ variable for fillers on target-present trials is not needed to compute innocent vs. guilty suspect discriminability across trials, but we compute it here anyway because, somewhat surprisingly, the mean of the difference score for fillers on target-present trials turns out to differ from the mean of the difference score for fillers on target-absent trials. The mean of x for a filler on target-present trials is equal to μ_{Lure} and the mean of y remains equal to (μ_{Target} + $\sum_{1}^{k-1} \mu_{Lure}$) / k. Thus, the mean of the $x - y$ variable for a filler in a target-present lineup is equal to μ_{Lure} - (μ_{Target} + $\sum_{1}^{k-1} \mu_{Lure}$) / k. Because μ_{Lure} = 0, the mean of this difference score reduces to 0 – (μ_{Target} + 0) / k = 0 – μ_{Target} / k = – μ_{Target} / k. Thus, according to this model, the mean of the filler distribution on target-present trials is actually shifted slightly below zero. This fact explains why the target-present filler distribution in Figure 5 differs slightly from the target-absent filler distribution.

We can use the expressions worked out above to specify the numerator of the d'_{IG} formula for the ability to discriminate innocent from guilty suspects across lineups according to the Ensemble model. More specifically, the numerator is equal to the mean of the $x - y$ decision variable for guilty suspects on target-present trials, which was found to be μ_{Target} - μ_{Target} / k above, minus the mean of the $x - y$ decision variable for innocent suspects on target-absent trials, which we determined is equal to 0. That is, the numerator of the d'_{IG} formula is equal to [μ_{Target} - μ_{Target} / k] – 0 = μ_{Target} - μ_{Target} / k = μ_{Target}(1 – 1/k).

The next goal is to compute the variance of the $x - y$ decision variable for the Ensemble model, which we denote Var($x - y$). When x corresponds to the guilty suspect on target-present trials,

$$\text{Var}(x) = \sigma^{2}_{Target}$$

$$\text{Var}(y) = \left(\sigma^2_{Target} + \sum_{k-1}\sigma^2_{Lure} + \sum_{i=1}^{k}\sum_{j\neq i}\rho\sigma_i\sigma_j\right)\Big/k^2$$

The expression for Var(y) is the general variance expression for the variance of the mean of k random variables, with all pairwise correlations equal to ρ. In Appendix A, we show that, in the equal-variance case (and with $\sigma^2 = 1$), the variance of the target-minus-ensemble decision variable is:

$$\text{Var}(x - y) = 1 - 1/k - [(k-1)/k]\rho$$

This variance expression is the same on target-present and target-absent trials in the equal-variance case, and the square root of that variance expression, $\sqrt{1-1/k-\rho(k-1/k)}$, is the denominator of the formula used to compute d'. Thus, for the Ensemble model, the ability to discriminate innocent from guilty suspects is given by:

$$d'_{IG} = \frac{\mu_{Target}(1-1/k)}{\sqrt{1-1/k-\rho(k-1/k)}}$$

When this result is rearranged into a simpler form (Appendix A), the Ensemble model predicts the following in the equal-variance case:

$$d'_{IG} = \frac{\mu_{Target}}{\sqrt{(1-\rho)\,k/(k-1)}} \tag{3}$$

According to this equation, as ρ increases, the denominator decreases. That is, the Ensemble model predicts that discriminability should increase as the correlation between memory strength signals increases.[4] Here again, we have assumed that the statistical computation – in this case, the computation of the difference between the memory signal for a given face and the average memory signal of all faces in the lineup – is an error-free process. Predicted discriminability would be lower than is implied by Equation 3 if we added random error to the computational process, but the predictions of the model with respect to ρ would not otherwise be affected.

For the equal-variance case involving a 6-person lineup (i.e., $k = 6$) and $\mu_{Target} = 2$ (the standard example we have used throughout), if $\rho = 0$, using Equation 3, the Ensemble model predicts that $d'_{IG} = 1.83$. This value is slightly less than the discriminability predicted by the Independent Observations model, which is $d'_{IG} = 2$ when $\mu_{Target} = 2$, regardless of the size of the correlation, but is greater than the discriminability predicted by the Integration model, which (as worked out above) is $d'_{IG} = 0.82$ when $\mu_{Target} = 2$ and $\rho = 0$. If, instead, $\rho = .50$, then, according to Equation 3, the Ensemble model predicts that $d' = 2.58$. This value is greater than the values of $d'_{IG} = 2.0$ and $d'_{IG} = 0.44$ for the Independent Observations model and Integration model, respectively, when $\mu_{Target} = 2$ and $\rho = .50$. Thus, in contrast to those models, the Ensemble model predicts that correlated memory signals should enhance the ability to discriminate innocent vs. guilty suspects. That prediction is illustrated in Figure 6A (based on a simulation in which, for the untransformed memory signals, $\mu_{Target} = 2$, $\mu_{Lure} = 0$, and $\sigma = 1$). Discriminability is obviously enhanced when memory signals are correlated whether all trials are considered (upper panel) or whether we consider only the subset of trials in which innocent or guilty suspects generated the MAX memory signal (Figure 6B).

[4] Mathematically, the Ensemble decision rule is linearly related to the Best – Rest model (Clark et al., 2011), where Rest equals the average of the other 5 lineup members rather than the ensemble average of all 6. These two models are linearly related and provide identical fits to the empirical data, so we do not distinguish between them.

Summary of model-based predictions about the effect of correlated memory signals

The key difference between the three models with respect to the role played by correlated memory signals is visually illustrated in Figure 7 using the simplest lineup scenario involving only two members. On target-present trials, the two members consist of the target (the guilty suspect) and a lure (a filler). On target-absent lineup trials, the two members consist of the replacement lure (the designated innocent suspect) and a lure (another filler). The figure shows the joint distribution of the suspect (*x*) and filler (*y*) memory strengths for target-present (black) and target-absent (grey) lineups. On target-present trials, $x \sim N(\mu_{Target}, \sigma_{Target})$ and $y \sim N(\mu_{Lure}, \sigma_{Lure})$, and on target-absent trials, $x \sim N(\mu_{Lure}, \sigma_{Lure})$ and $y \sim N(\mu_{Lure}, \sigma_{Lure})$. As illustrated in the figure, where the correlation between *x* and *y* is set to .80, the different decision rules can be thought of as collapsing these joint distributions in different ways. The independent observations decision variable for guilty and innocent suspects amounts to the distribution along *x* for target-present and target-absent trials (which is unaffected by the presence of a positive correlation). By contrast, in the presence of a positive correlation, the Integration (additive) variable increases variance and yields lower separation than the independent decision variable. The Ensemble (difference) variable instead yields reduced variance and, therefore, a greater separation between guilty and innocent suspects than the "independent" target signal alone.

Likelihood Functions

Thus far, we have illustrated the predictions that each of the three models makes about the effect of correlated memory signals on the underlying memory-strength distributions. Doing so was relatively straightforward, but specifying their corresponding likelihood functions is more challenging. The likelihood functions are needed to fit the models to empirical data. They go beyond the equations presented thus far in that they specify the probability of a suspect ID, a filler ID or no ID on a given target-present or target-absent trial. We first describe a general conceptualization of correlated memory signals in terms of shared vs. unshared variance that greatly facilitates our subsequent derivation of the model-specific likelihood functions.

Shared Variance (Correlated Memory Signals)

Partitioning Target and Lure Distribution Variance. Consider the 5 X 5 matrix of target and lure distributions shown in Figure 8. The distributions depict the raw (untransformed) memory signals that are used by the Independent Observations model, and the bottom row depicts 5 identical signal detection scenarios that directly correspond to the model illustrated in Figure 1. That is, for the scenarios in the bottom row of Figure 8, the memory strength values for lures (innocent suspects and fillers) and for targets (guilty suspects) are distributed according to Gaussian distributions with means of μ_{Lure} and μ_{Target}, respectively, and, under the equal variance assumption (i.e., $\sigma_{Target} = \sigma_{Lure}$), the same standard deviation, denoted here as σ. Thus, the ability to discriminate innocent from guilty suspects is given by $d'_{IG} = (\mu_{Target} - \mu_{Lure}) / \sigma$.

Figure 8 illustrates the fact that the variance of the memory signals for innocent and guilty suspects in all 5 models in the bottom row (with variances for both distributions fixed at σ^2) can arise from different sources. More specifically, σ^2 can be partitioned into the variance of the mean memory signal *between* lineups (which we denote σ^2_b) and the variance of the individual item memory signals *within* a lineup (which we denote σ^2_w) such that $\sigma^2 = \sigma^2_b + \sigma^2_w$. The situation is exactly analogous to a one-way repeated-measures ANOVA, with σ^2_b corresponding to between-subject variance and σ^2_w corresponding to within-subject variance. In formal terms, shared variance is distributed as $b \sim N(0, \sigma_b)$, and the targets (x) and lures (y) are distributed as $x \sim b + N(\mu_{Target}, \sigma_w)$, and $y \sim b + N(\mu_{Lure}, \sigma_w)$. Thus, the means of the target and lure distributions would be μ_{Target} and μ_{Lure}, respectively, and their corresponding (equal) variances would be $\sigma^2_{Target} = \sigma^2_b + \sigma^2_w$, and $\sigma^2_{Lure} = \sigma^2_b + \sigma^2_w$.

Rows 1 through 4 in Figure 8 show hypothetical distributions from which memory signals are drawn for four separate lineups (lineup 1 through lineup 4, as labeled on the right side of the figure). In column 1, for all four individual lineups, the memory signal for the guilty suspect in a target-present lineup is drawn from the same target distribution (with mean μ_{Target} and standard deviation $\sigma_{Target} = \sigma$). Similarly, the memory signals for the innocent suspect in a target-absent lineup and for the fillers in both target-present and target-absent lineups are drawn from the same lure distribution (with mean μ_{Lure} and standard deviation $\sigma_{Lure} = \sigma$). This column illustrates the simplest case, where the correlation of memory signals for items within a lineup is zero. The correlation is zero because the memory signal for the suspect in a particular lineup does not predict the memory signals of the fillers in that lineup. A special feature of the 0-correlation scenario is that there is no variance in the mean memory signal *between* lineups (i.e., $\sigma^2_b = 0$), so all of the variance in the aggregate target and lure distributions shown in the bottom row of column 1 (σ^2) comes from the variance of the item memory signals within a lineup (σ^2_w). That is, because $\sigma^2 = \sigma^2_b + \sigma^2_w$, under the 0-correlation scenario where $\sigma^2_b = 0$, $\sigma^2 = 0 + \sigma^2_w = \sigma^2_w$. Thus, $d'_{IG} = (\mu_{Target} - \mu_{Lure}) / \sigma_w$. The state of affairs illustrated in column 1 corresponds to how signal detection theory has been used to conceptualize lineup performance in the past in terms of the Independent Observations model.

In reality, we assume that the means of the target and lure distributions are likely to differ across lineups (i.e., σ^2_b is likely to be greater than 0). As noted earlier, their means are likely to differ because the police create lineups not by randomly selecting faces but by instead selecting faces that correspond to the description of the perpetrator, thereby ensuring that the faces share features. Columns 2 through 5 illustrate varying degrees of across-lineup variance of mean memory signals, which is visually evident in the fact that the means of the distributions from which the target and lure memory signals are drawn now vary from lineup to lineup. In other words, now, $\sigma^2_b > 0$, and its value increases from column 2 to column 5. It is also visually apparent that, as the variance of mean memory signals across lineups (σ^2_b) increases, the variance of the memory signals of items within a lineup (σ^2_w) must decrease to maintain the same signal detection scenario across trials (shown in the bottom row) in which target and lure variance are both fixed at σ^2. In other words, because $\sigma^2 = \sigma^2_b + \sigma^2_w$, and because the value of σ^2 in the bottom row is fixed, as σ^2_b increases, σ^2_w decreases.

This non-zero variation of mean memory signals across lineups implies that some of the variance in the aggregate memory signals shown in the bottom row of Figure 8 is shared by the faces in a given lineup. This shared variance means that the strength of the memory signal associated with the suspect in any given lineup is correlated with (i.e., is predictive of) the strength of the memory signals associated with the fillers in that lineup. The magnitude of the correlation, ρ, is equal to the ratio of the shared variance (σ^2_b) to the total variance ($\sigma^2_b + \sigma^2_w$). That is, $\rho = \sigma^2_b / (\sigma^2_b + \sigma^2_w)$. This is the same formula that has been used to calculate the intraclass correlation coefficient when assessing interrater reliability for a random sample of *n* judges rating a set of *k* target items (e.g., Case 2 in Shrout & Fleiss, 1979).

As noted above, column 1 of Figure 8 shows one extreme in which all of the aggregate variance in the bottom row arises from within-lineup variance. In that case, $\sigma^2_b = 0$, so $\sigma^2 = \sigma^2_w$ and $\rho = 0 / (0 + \sigma^2_w) = 0$. In contrast, column 5 shows the opposite extreme in which all of the variance in the aggregate distributions arises from across-lineup variance. In that case, $\sigma^2_w = 0$, so $\sigma^2 = \sigma^2_b$, and $\rho = \sigma^2_b / (\sigma^2_b + 0) = 1$. Yet in all cases, the ability to discriminate innocent from guilty suspects, as depicted in the bottom row of each column, is given by d'_{IG} = (μ_{Target} - μ_{Lure}) / √(σ^2_b + σ^2_w), with σ^2_b + σ^2_w equal to the fixed value σ^2.

Within-lineup discriminability. The ability to discriminate innocent from guilty suspects (d'_{IG}) is an inherently across-lineup measure because a given lineup contains either an innocent suspect or a guilty suspect (not both). Nevertheless, it is also of interest to consider *within-lineup d'* for target-present lineups because its value changes as a function of ρ even when d'_{IG} is held constant (as it is in the bottom row of Figure 8). The ability to discriminate the guilty suspect from the fillers in a target-present lineup is given by d'_{TP} = (μ_{Target} - μ_{Lure}) / σ_w. This *d'* formula applies to all of the signal detection models depicted in Figure 8 except for the net (aggregate) distributions presented in the bottom row, where it is always the case that $d'_{IG} = \left(\mu_{Target} - \mu_{Lure}\right)\big/\sqrt{\sigma_b^2 + \sigma_w^2}$. In the 0-correlation scenario where σ^2_b = 0 (column 1), d'_{TP} is equal to d'_{IG}. However, as σ^2_b increases from column 1 to column 5, σ^2_w decreases. As the correlation approaches 1 (such that σ_w approaches 0), d'_{TP} approaches infinity, in which case the guilty suspect could be correctly picked out of the lineup every time. This is true even though the ability to discriminate innocent from guilty suspects (d'_{IG}) would be unaffected.

As noted earlier, this beneficial effect of correlated memory signals for target-present lineups (namely, $d'_{TP} \rightarrow \infty$ as $\rho \rightarrow 1$) is the same beneficial effect of correlated memory signals that is observed in the 2AFC task. In contrast to target-present lineups, the ability to discriminate the innocent suspect from fillers for fair *target-absent* lineups (d'_{TA}) is, by definition, always equal to 0 because, for fair lineups, the innocent suspect is just another filler from the witness's point of view. For that reason, the size of the correlation does not affect the ability to discriminate the innocent suspect from fillers within fair target-absent lineups. Thus, the chances that an ID would land on the innocent suspect is 1/6 for a fair 6-member lineup despite the reduction in within-lineup variance as the correlation increases.

Differentiation. For the models we have considered thus far, an issue that could complicate the interpretation of correlated memory signals depicted in Figure 8 is that signal detection models for list-memory designs are often assumed to rely on a likelihood ratio decision rule (Glanzer & Adams, 1985; McClelland & Chappell, 1998; Shiffrin & Steyvers, 1997). In the likelihood ratio version of these models, the decision would instead be based on the likelihood that the test item was drawn from the target distribution divided by the likelihood that it was drawn from the lure distribution (e.g., Semmler, Dunn, Mickes & Wixted, 2018). The potential complication is that likelihood ratio models inherently predict a phenomenon known as *differentiation*.

When targets and lures share few features (unlike in the lineup situation), differentiation results in the target and lure distributions moving in opposite directions (not in the same direction, as they do in rows 1 through 4 of Figure 8). When the targets and lures share many features, as they presumably do in a well-constructed lineup, likelihood ratio models instead predict that as the target distribution shifts to the right, the lure distribution also shifts to the right but to a lesser degree, thereby increasing the separation of the two distributions (see, for example, Figure 2 of Criss & McClelland, 2006). Thus, differentiation would still be observed in that sense. We mention this because the target and lure distributions shown in Figure 8 differ from that pattern in that they shift in lockstep.

One way of conceptualizing the differentiation scenario would be to assume that targets (x) and lures (y) are distributed as $x \sim b + N(\mu_{Target}, \sigma_w)$ and $y \sim \lambda b + N(\mu_{Lure}, \sigma_w)$, where $0 < \lambda < 1$. Thus, shared variance would cause the lure distribution to shift to a lesser degree than the target distribution, as would be true in the differentiation scenario. Consider the extreme case in which all of the variance in the net distributions arises from shared variance (i.e., $\sigma_w = 0$, right column of Figure 8). In that case, the standard deviations of the net target and lure distributions would be equal to σ_b and $\lambda\sigma_b$, respectively. In other words, an unequal-variance model would be expected, with the variance of the target distribution exceeding the variance of the lure distribution ($\sigma_{Target} > \sigma_{Lure}$), which is the pattern typically observed in list-memory studies (Egan, 1958; Wixted, 2007). However, in our later fits of the models to empirical lineup data, and in our prior model-fitting studies (e.g., Wixted, Mickes, Dunn, Clark & Wells, 2016), we have never observed that pattern. Instead, we either find that $\sigma_{Target} = \sigma_{Lure}$ or, for understandable reasons considered in more detail later, $\sigma_{Target} < \sigma_{Lure}$. Thus, we use the memory-strength distributions shown in Figure 8 in our model-specific likelihood function derivations and assume that any differentiation that might exist is small enough that it can be ignored.

Although we assume that a likelihood ratio decision rule is not applied to the raw memory-strength signals illustrated in Figure 8, our analysis is still fully compatible with a likelihood ratio decision rule applied to the additive diagnostic variable of the Integration model or to the subtractive diagnostic variable of the Ensemble model. Only the Independent Observations model, for which the raw memory-strength signal is the diagnostic variable, would be excluded from a likelihood ratio interpretation. In any case, we use the lockstep interpretation of correlated raw memory signals presented in Figure 8 to facilitate the derivation of the model-specific likelihood functions described next.

Model-Specific Likelihood Functions

The derivation of the relevant likelihood functions begins by specifying the joint probabilities of the "events" that result in a given outcome for a particular face (i.e., an outcome consisting of an ID with a particular level of confidence or no ID). The events are as follows: (1) the probability of observing a given memory strength, x_i, for the face in question, (2) the probability that x_i is the MAX value in the lineup, and (3) the probability that the decision variable, $f(\mathbf{x})$, exceeds the decision criterion for making an ID with a particular level of confidence, where **x** is the set of all items in a given lineup. That is, $\mathbf{x} = \{x_1, x_2, x_3, \ldots x_k\}$, where k is lineup size. When all three conditions are satisfied, the face is identified with the level of confidence corresponding to the highest confidence criterion exceeded by $f(\mathbf{x})$. If the three conditions are not satisfied by any face in the lineup, then no ID is made (i.e., the lineup is rejected). Events 1 and 2 are the same for all three models, but the models differ with respect to event 3. That is, they differ with respect to the decision variable, $f(\mathbf{x})$.

As an example, consider the probability of identifying the guilty target with memory strength x_1 from a target-present lineup. There is (1) some probability of observing a particular memory strength of the target, x_1, (2) some probability that x_1 will be the highest (MAX) memory strength of the lineup members, and (3) some probability, $f(\mathbf{x})$, that the decision variable will exceed the decision criterion. The joint probability of those events is the probability that the target will be identified from a target-present lineup. For the Independent Observations model, the decision variable, $f(\mathbf{x})$, is x_1 itself. For the Integration model, $f(\mathbf{x}) = \sum_{j=1}^{k} x_j$, where x_j represents the memory strength of the jth face in the lineup. For the Ensemble model, $f(\mathbf{x}) = x_1 - 1/k \sum_{j=1}^{k} x_j$.

Assuming a standard signal detection model, the probability of observing target memory strength x_1 (event 1) is given by a Gaussian distribution with mean, $\mu_1 = \mu_{Target}$ and variance $\sigma_1^2 = \sigma^2_{Target}$:

$$P(x_1) \;= \frac{1}{\sqrt{2\pi\sigma_1^2}}\, e^{-(x_1-\mu_1)^2/(2\sigma_1^2)} \tag{4}$$

The probability that x_1 is greater than the memory strength of a particular filler *j* is obtained by integrating a Gaussian distribution with mean $\mu_j = \mu_{Lure}$ and variance $\sigma_j^2 = \sigma^2_{Lure}$ from -∞ to x_1:

$$\frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{x1} e^{-(x_1-\mu_j)^2/(2\sigma_j^2)}dx_j \;=\; \Phi\left(\frac{x_1-\mu_j}{\sigma_j}\right)$$

where Φ is the standard cumulative normal distribution. Thus, the probability that a given x_1 is greater than the value of *all* fillers in a lineup of size *k* (event 2) is:

$$P(x_2 \ldots k < x_1|x_1) = \prod_{j=2}^{k} \Phi\left(\frac{x_1-\mu_j}{\sigma_j}\right) \tag{5}$$

And the probability that the decision variable, *f*(**x**), exceeds the decision criterion, *c*, given x_1 (event 3) is simply:

$$P(f(\mathbf{x}) > c \mid x_1) \tag{6}$$

where, again, x_1 is the memory strength of the target in this example. Thus, the probability of observing x_1 *and* the probability that x_1 is greater than the value of all lures in a lineup of size *k* *and* the probability that the decision variable, *f*(**x**), exceeds the decision criterion, integrated over all possible values of x_1 (i.e., over all possible target memory-strength values) is given by Equation 4 × Equation 5 × Equation 6 integrated from -∞ to +∞:

$$\frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{+\infty} P(x_1)\,P(x_2 \ldots x_k < x_1|x_1)P(f(\mathbf{x}) > c\;|x_1, x_2 \ldots x_k < x_1)dx_1$$

Or, in more detail,

$$\frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{+\infty} e^{-(x_1-\mu_1)^2/(2\sigma_1^2)}\;\prod_{j=2}^{k}\Phi\left(\frac{x_1-\mu_j}{\sigma_j}\right)\;P(f(\mathbf{x}) > c\;|x_1, x_2 \ldots x_k < x_1)dx_1$$

Again, this is the likelihood of observing a *target* (i.e., guilty suspect) ID from a target-present lineup. Similar equations express the probability of observing a filler ID or a no ID from target-present and target-absent lineups. The 3 models yield different estimates for each probability because $f(\mathbf{x})$ differs for each model. The full details for each of these likelihood functions (separately for the Independent Observations, Integration and Ensemble models, for both target-present and target-absent lineups) are presented in Appendix B.

For both the Integration and Ensemble models, the derivation provided in Appendix B involves a Gaussian approximation of a variable that is not truly Gaussian. An approximation is required because, for example, for a given value of x_1 associated with a target when it is the MAX value in the lineup, the distribution of memory strengths for the fillers are drawn from a truncated Gaussian distribution that ranges from $-\infty$ to x_1. Thus, the expected value of the mean of those 5 fillers given x_1 (and given that x_1 is the MAX value) involves computing the sum or mean of 5 non-Gaussian variables. Our derivation relies on the assumption that this aggregated value is distributed as a Gaussian variable. According to the central limit theorem, that would be a safe assumption when many variables are averaged, but with only 5 variables summed or averaged, the distribution would not necessarily be approximately Gaussian. Nevertheless, even under these conditions, we found that the Gaussian approximation is extremely accurate and does not detectably affect the ability to distinguish between the competing models (see section entitled "Truncated normal approximation" in Appendix B).

Model Parameters and Model-Recovery Simulations

We next describe the specific parameters to be estimated for each model and then report model-recovery simulations in which simulated data were generated for each of the three models. For each simulated dataset, all three models were fit to the data to determine (1) which model fit best (it should be the model that generated the simulated data) and (2) whether the estimated parameters corresponded to the programmed parameters.

Model Parameters

In the models we later fit to empirical data, we set $\mu_{Lure} = 0$ and $\sigma_{Lure} = 1$ and estimate the following free parameters: (1) μ_{Target}, (2) σ_{Target}, (3) σ^2_b, and (4) the confidence criteria (e.g., c_1 through c_n when an *n*-point confidence scale is used). For the equal-variance signal detection model, $\sigma^2_{Target} = \sigma^2_{Lure} = \sigma^2$. As noted earlier (e.g., Figure 8), in that case, $\sigma^2 = \sigma^2_b + \sigma^2_w$. Because we set σ^2 equal to 1 for convenience, it follows that $\sigma^2_w = 1 - \sigma^2_b$. Thus, estimating σ^2_b from a fit to correlated data automatically estimates σ^2_w as well, so only one parameter (σ^2_b) is needed to estimate the correlation, where, again, $\rho = \sigma^2_b / (\sigma^2_b + \sigma^2_w) = \sigma^2_b / (\sigma^2_b + 1 - \sigma^2_b) = \sigma^2_b$. In other words, σ^2_b is the correlation parameter. As described in Appendix B, the Independent Observations model and the Integration model both require this parameter to capture correlated memory signals, but the Ensemble model does not because it subtracts out shared variance regardless of the size of the correlation. Analogously, the computational formula used to compute *d'* for 2AFC recognition does not require a correlation parameter even though, for that task (as we described earlier), $d' = (\sqrt{2})\ \mu_{Target} / (1 - \rho)$. Whether the memory signals are correlated or uncorrelated, the computational formula for the 2AFC task is $(1/\sqrt{2})[z(H) – z(F)]$, where *H* and *F* represent the hit and false alarm rates (see Equation 7.2 of Macmillan & Creelman, 2005, p. 372). This formula does not require ρ as a parameter because it assumes a subtractive decision rule, which means that shared variance is subtracted out. For the same reason, the Ensemble model does not require σ^2_b as a free parameter even when the memory signals in a lineup are correlated. Thus, typically, the Ensemble model requires one fewer free parameter than the other two models.

All of the models include σ_{Target} as a free parameter to allow for the possibility of unequal target and lure variances. In list-memory studies, the target and lure distributions have usually been found to have unequal variances, with the targets having greater variance than the lures (Egan, 1958; Wixted, 2007). One explanation for that finding is that variable amounts of memory strength are added to the target items during encoding (e.g., due to random variability in a subject's attention across the study list). If that were true, then both the mean and the variance of the targets would increase relative to the lures. The same phenomenon might be expected to increase the variance of the target distribution relative to the lure distribution when memory is tested using a lineup. Then again, as shared variance (σ^2_b) increases, any effect of encoding variability on the target distribution would increasingly apply to the lure distribution as well (counteracting the differential effect of encoding variability on targets). Thus, an unequal-variance model might be less likely to be observed in a lineup study compared to a list-memory study. Indeed, in none of the fits described later is $\sigma_{Target} > \sigma_{Lure}$. As noted earlier, this empirical result is one reason why we do not assume that the distributions in Figure 8 exhibit differentiation (and instead assume that they shift more-or-less in lockstep).

Another variable that can affect the relative variances of the target and lure distributions in a lineup study is the size of the pool of the stimuli from which the faces in the lineup are drawn. In some of the experiments we will consider later, the lures are randomly drawn from a large pool of faces (different lures for different subjects), whereas the same target face is used for every subject (namely, the one face that matches the perpetrator seen in the video). A design like that would be expected to selectively add variability to the memory strengths of the lures. For example, by chance, some of the lures might look very much unlike the perpetrator, but some others might be virtual lookalikes.

The key point is that σ^2_w can be partitioned into multiple sources of variance that can differentially affect the variance of the target distribution relative to the lure distribution. To allow for differential effects of these sources of variance, we can define $\sigma_{w\text{-}target} = \alpha\sigma_w$, where σ^2_w now specifically refers to the within-lineup variance of the lures. Thus, in the general (non-equal-variance) case, targets (x) and lures (y) are distributed as $x \sim b + N(\mu_{Target}, \alpha\sigma_w)$, and $y \sim b + N(\mu_{Lure}, \sigma_w)$. Recall that for the equal-variance case, the variances of the aggregate target and lure distributions are both equal to $\sigma^2_b + \sigma^2_w$. For the unequal-variance we are considering now, $\sigma^2_{Lure} = \sigma^2_b + \sigma^2_w$ but $\sigma^2_{Target} = \sigma^2_b + \sigma^2_{w\text{-}target} = \sigma^2_b + \alpha\sigma^2_w$. Thus, estimating σ_{Target} provides an indirect estimate of α.

Model-Recovery Simulations

To confirm the validity of the likelihood functions derived in Appendix B, we conducted extensive model-recovery simulations. These simulations were conducted to ensure that the full set of simulated data generated by a particular model (guilty suspect IDs, filler IDs, and no IDs from target-present lineups, plus filler IDs and no IDs from target-absent lineups) would be accurately fit by that model while at the same time uniquely recovering the programmed parameter values. In each of many model-recovery simulations, three sets of simulated data were first created using a given set of parameter values, one using the Independent Observations model, one using the Integration model, and one using the Ensemble model. Each simulated data set was then fit with the Independent Observations, Integration, and Ensemble models using the likelihood functions derived in Appendix B (maximizing the likelihood of the data). Thus, there were 9 fits per round of model-recovery simulations.

When generating simulated data for a given model, there were always 6 members in a lineup, and a 6-point confidence scale was always used. We set $\mu_{Lure} = 0$ and $\sigma_{Lure} = 1$, and we set

c_1 through c_6 to fixed values that differed from the three models (these values were chosen so that the predicted ROC data would fall in a reasonable range). Although we conducted many sets of model-recovery simulations, for the model-recovery results presented here, we set $\mu_{Target} = 1.5$ and $\sigma_{Target} = 1$. The simulated data were generated with shared variance distributed as $b \sim N(0, \sigma_b)$ and with lures (x) and targets (y) distributed as $x \sim b + N(\mu_{Lure}, \sigma_w)$, and $y \sim b + N(\mu_{Target}, \sigma_w)$. For one set of simulations involving uncorrelated data, we set $\sigma^2_b = 0$, whereas for a second set of simulations involving correlated data, we set $\sigma^2_b = .50$.

Figure 9 shows the χ^2 goodness-of-fit results for the uncorrelated and correlated model recovery simulations we performed, with details presented in Table 2. For a given simulation, there were 10,000 target-present trials and 10,000 target-absent trials. Obviously, the chi-square goodness-of-fit statistics indicate that, in each case, the model that generated the data also fit the data extremely well and fit better than the two alternative models. In two cases, the fit of the wrong model appears to at least rival the fit of the true model. Those two cases involve the correlated lineup data generated by either the Independent Observation model or the Integration model (Figure 9B). Although both models can fit the data generated by the other model fairly well, the apparent rivalry is an illusion because, in both cases, the correct model (but not the incorrect model) could fit the data with one fewer free parameter by fixing $\sigma_{Target} = 1$. We allowed that parameter to vary to ensure that the correct model would recover its true value, which it always did (Table 2).

Indeed, of more importance than the goodness-of-fit data is the fact that, for both the uncorrelated and correlated simulated data, all of the programmed parameter values (not just σ_{Target}) were recovered with almost perfect accuracy for the 6 cases in which a model was fit to its own simulated data. This can be seen by comparing the parameter estimates in Table 2 shown in bold relative to the “true” programmed values shown in italics. The true parameters were accurately recovered in all of our model-recovery simulations, not just those shown in Table 2. The results of these simulations demonstrate that the likelihood functions derived in Appendix B are accurate even though, for the Integration and Ensemble models, they involve a very close Gaussian approximation of a non-Gaussian random variable. Note that the Ensemble model has a particularly hard time fitting data generated by the other two models (Figure 9). These results may indicate that the Ensemble model is the least flexible of the three models under consideration.

One final source of variance might sometimes need to be estimated as well, namely, criterion variance (Benjamin, Diaz & Wee, 2009). Confidence criteria surely vary across participants. For the Independent Observations and Integration models, the same parameter that captures shared variance ($\sigma^2{}_b$) also captures criterion variability. In other words, although we have conceptualized $\sigma^2{}_b$ as the variance in the means across lineups (shared variance), for these models, it can instead be conceptualized as the variance of confidence criteria shifting in lockstep across lineups (or as a combination of the two sources of variance). For the Ensemble model, by contrast, an additional parameter (σ_c) would be needed to capture lockstep criterion variance. For the fits we performed to empirical data (described in the next section), adding a criterion-variance parameter to the Ensemble model never improved its fit. However, it is conceivable that this parameter would need to be added to the Ensemble model for it to adequately fit other data sets collected in future studies.

Our approach to capturing criterion variance does not take into account the possibility of independent criterion noise over and above a lockstep shift across trials, and it is not entirely clear to us how to write the likelihood functions in such a way as to capture that additional source of variance. We therefore investigated the effect of independent criterion noise by repeating the above simulations except that independent criterion noise was added to the simulated data. This was accomplished by selecting each confidence criterion from a distribution, such that $c_i \sim N(\mu_i, \sigma_v)$, where μ_i represents the mean placement for confidence criterion *i* and $\sigma_v = 0.50$. The only restriction was that the confidence criteria remain monotonically arranged. Table 3 shows that the model-fitting pattern remains largely unchanged despite the presence of independent criterion noise. The programmed parameter values are recovered somewhat less accurately, and the overall fits are not quite as good for the Ensemble model (which also needed its lockstep criterion-variance parameter, σ_c, to provide an adequate fit), but these results suggest that moderate independent criterion variability should not dramatically affect the interpretation of which model best accounts for the empirical data. Here again, note that the fit of the correct model (but not the incorrect model) would remain largely unchanged if we fixed $\sigma_{Target} = 1$ instead of allowing to vary as a free parameter. Thus, the advantage of the correct models over the incorrect models in Table 3 is larger than the χ^2 values imply when taken at face value.

Finally, one potentially problematic issue came to light in our model recovery simulations. To increase ecological validity, eyewitness identification researchers often use several different targets (i.e., several different perpetrators) in a study instead of having all subjects watch the same mock-crime video. In these multiple-target studies, each target is presented to a different subset of the subjects in the study, and each target has their own description-matched lineups. Basically, each target is used in its own mini lineup study, and then the data are pooled together for the final large-*N* analysis. The potential problem for model-fitting purposes is that different targets can, and usually do, give rise to different d'_{IG} values (because, for example, some targets are more memorable than others, exposure time is greater for some targets than others, etc.). Our model-recovery simulations showed that when the data are pooled across multiple targets who are associated with different d'_{IG} values (creating a mixture model), the model that generated the data is not necessarily the model that provides the best fit. Therefore, for purposes of identifying theoretical mechanisms, we focus on large-*N* studies that used single-perpetrator designs.

Model Fits to Empirical Data

Having derived and validated their likelihood functions, the next step is to fit the three competing models to empirical data. The goal of model-fitting is not so much to identify the winning model as it is to rule out models that are not viable. As noted by Pashler and Roberts (2000), the mere fact that a model provides a good fit cannot be assumed to validate that model. However, a model that provides a differentially poor fit relative to other models can be reasonably rejected. Although our main focus here is obviously on model fitting, we also later review the relevant non-model-fitting evidence bearing on the predictions of the three competing models.

We fit each of the three models to several empirical data sets. The data were taken from eyewitness identification experiments in which (1) a large number of subjects (~1000) were tested in a given condition, (2) all of the subjects viewed the same target, (3) the subjects were tested only once, and (4) the lineups were fair. Several studies fit our criteria. Recently, for example, Mickes et al. (2017) reported data from simultaneous lineups presented to a large number of subjects, and they collected ROC data in two different ways. One ROC was created using confidence ratings (the typical approach) and the other was created using an instructional biasing manipulation. Thus, this data set is unique in that the model-fitting results can be tested for generality across different methods for generating the ROC data. Subjects in this experiment were randomly assigned either to a confidence rating condition (*N* = 978) or to one of four instructional biasing conditions: liberal (*N* = 1066), neutral (*N* = 1037), unbiased (*N* = 984), or conservative (*N* = 1076). In each condition, approximately half the subjects were randomly assigned to a target-present lineup and half to a target-absent lineup. One ROC was constructed using data from the confidence rating condition, and the other was constructed using data from the four instructional biasing conditions (i.e., the instruction-based ROC had 4 points, one for each biasing condition).

Figure 10 shows the empirical ROC data from this experiment. The ROC data from the confidence condition are shown as filled gray circles. The solid black curve represents an atheoretical fit provided by pROC software (with estimated standard errors of the fit shown in light gray). This software package is often used to compute (atheoretical) partial area under the ROC curve. The four open symbols represent the correct and false ID rates from the four different biasing conditions (upright triangle = liberal instructions, inverted triangle = neutral instructions, circle = unbiased instructions, and square = conservative instructions). The dashed diagonal line represents chance performance. The top horizontal axis (TA Filler ID Rate) is included as a reminder that the false ID rate shown on the bottom axis is the TA filler ID rate divided by lineup size because the lineups were fair and there was no designated innocent suspect in this study. Thus, the estimated false ID rate for the innocent suspect is equal to the TA filler ID rate divided by the lineup size of *k*. Note that the two extreme biasing conditions yield points that appear to fall on a slightly lower ROC curve compared to the two more neutral biasing conditions and compared to the confidence ROC curve.

We first fit the Independent Observations, Ensemble, and Integration models to the confidence-based ROC data using the relevant likelihood functions described earlier. We optimized the fits by maximizing the likelihood of the data. Note that these models were fit to the full data set (including filler IDs from target-present lineups), not just to the subset of ROC data shown in Figure 10 (which does not represent target-present filler IDs because the ROC represents suspect IDs). Table 4 shows the results of the model fits. The Ensemble model fit the data better (i.e., it yielded a lower chi-square) than the two competing models, with the Integration model performing the worst. Figure 11 shows the observed and predicted data. All of the models capture the trends in the standard ROC data (Figure 11A), but an advantage for the Ensemble model is apparent for the target-present ROC data, which plots the target-present suspect ID rate vs. the target-present filler ID rate (Figure 11B). For these data, the Ensemble model provided a closer approximation than the competing models, perhaps because the Ensemble model uniquely predicts that the mean memory strength of target-present fillers differs from that of target-absent fillers (see Figure 5).

The performance of the Ensemble model in this case is even better than it might seem to be at first glance. Given that the data likely involved correlated memory strength signals, it is perhaps not surprising that the Independent Observations model needed the σ^2_b parameter to adequtely fit the data. But even with that extra parameter, it still yielded a higher chi-square than the Ensemble model. Thus, according to AIC and BIC, which penalize models for having extra free parameters, the Ensemble model is also (necessarily in this case) judged to have provided the best fit. The Integration model provided a particularly poor fit in that the observed data deviated significantly from its optimal predictions. Adding the σ^2_b parameter (to capture correlated memory-strength values) did not significantly improve its fit. Thus, these data weigh against the Integration model and in favor of the Ensemble model, but they do not necessarily reject the Independent Observations model (the predictions of which did not deviate significantly from the observed data).

The goodness-of-fit story is similar for the instruction-based ROC data as shown in Table 5. These data were actually fit twice by all three models, once assuming a single value of μ_{Target} and once again allowing μ_{Target} to differ for the two extreme biasing conditions compared to the two more neutral conditions. For all three models, the fit was significantly improved by allowing μ_{Target} to differ in this way ($\mu_{Target1}$ corresponds to the two extreme biasing conditions, and $\mu_{Target2}$ corresponds to the two neutral biasing conditions), but the relative standing of the three models was unchanged. As shown in Table 5, the Ensemble model once again clearly provided the best fit according to all of the goodness-of-fit measures (χ^2, AIC and BIC). The Independent Observations model, even with an extra free parameter ($\sigma^2{}_b$), did not yield a chi-square value as low as the Ensemble model did. Moreover, the deviations between predicted and observed data for the Independent Observations were significant in this case, though not by much. And once again, the Integration model provided the poorest fit (deviating significantly from the data), one that was not significantly improved by adding the $\sigma^2{}_b$ parameter to capture correlated memory-strength values. Figure 12 shows the observed and predicted data. Once again, all of the models capture the trends in the standard ROC data (Figure 12A), but a slight visual advantage is apparent for the Ensemble model for the target-present ROC data (Figure 12B).

For both the confidence-based ROC data and the instruction-based ROC data, an unequal-variance model is suggested by the Ensemble model, with the standard deviation of the target distribution estimated to be *less* than that of the lure distribution in both cases (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). As noted earlier, one possibility is that this result 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 lure distribution with greater variance than the target distribution (as suggested by the best-fitting Ensemble model).

Next, we compared the ability of the three models to fit the data from a second large-*N* study. Seale-Carlisle & Mickes (2016) compared the simultaneous lineup to another kind of sequential lineup used in the UK (we refer to this as the US vs. UK study). In US procedure (*N* = 1148), 6 faces were shown simultaneously. In the UK procedure (*N* = 1057), 9 faces were presented sequentially, and each face was presented as a moving video instead of as a still photo. Unlike the standard sequential procedure used in the US, in the UK procedure, witnesses lap through the 9 faces twice before making a decision. Thus, in principle, they could make a memory-based comparison between the best face vs. the ensemble of the full set of faces in the lineup, in which case the Ensemble model might be the most appropriate model for both lineup procedures.

The stimuli used in the US vs. UK study were completely different from the stimuli used for the confidence-based vs. instruction-based ROC study described above. The witnessed event consisted of a 20-s mock-crime video of a young White male stealing several items from a vacated office. An experienced London Metropolitan Police Officer with specialized training in eyewitness identification procedures filmed the actor according to legally mandated specifications from the Police and Criminal Evidence (PACE) act of 1984 (Code D). The Officer also selected nine fillers based on PACE code guidelines from the database used by the London Metropolitan Police Force for constructing lineups. Thus, in this study, the fillers were not selected randomly from a large pool of stimuli for each subject. Instead, the stimuli were fixed for subjects assigned to target-present lineups (containing the guilty suspect) and for subjects assigned to target-absent lineups (containing a replacement filler). Using a fixed set of stimuli across subjects is potentially methodologically problematic in terms of ecological validity. However, on the positive side, the fact that the stimuli were selected by an experienced police officer presumably works in the opposite direction, enhancing ecological validity. In any event, the results of the model fits turned out to be similar to the fits described above and are presented in Table 6.

For both the US and UK procedures, the Ensemble model, despite having the fewest free parameters, again provided the best fit according to all goodness-of-fit measures. An equal-variance model was implied by the Independent Observations and Ensemble models (i.e., allowing for unequal variance did not significantly improve the fit for either model).[5] According to AIC and BIC, of the three models, the Integration model provided the worst fit to the US data and the second worse fit to the UK data. In addition, its best-fitting parameter estimates for the UK data were slightly odd, with μ_{Target} estimated to be 0.

Note that, according to all three models, the simultaneous (US) procedure far outperformed the sequential (UK) procedure in terms of discriminability. For example, according to the best-fitting Ensemble model, d'_{IG} for the US procedure was 1.27, whereas d'_{IG} for the UK procedure was 0.60 (for these equal-variance fits, $d'_{IG} = \mu_{Target}$). The two lineup procedures differ

[5] This equal-variance finding may reflect the fact that, in this study, item variance for fillers was minimized because the fillers were not randomly drawn from a large group of photos.

in many ways, and it is not clear which differences account for the discriminability advantage enjoyed by the simultaneous US lineups. One possibility is that an ensemble representation based on memory for sequentially presented faces is noisier than an ensemble representation based on faces that are simultaneously available. Whatever the reason for the difference in discriminability, the key point for our purposes here is that the Ensemble model fit the data from both procedures the best even though the competing models had an additional free parameter.

Finally, we fit data from another large-*N* single-target study reported by Brewer and Wells (2006). In this study, subjects first watched a video in which they viewed two targets, a thief and a waiter. All 1200 subjects were then tested for their ability to identify the thief from an 8-member simultaneous lineup (with the stimuli fixed across subjects for both target-present and target-absent lineups). Thus, this was a single-target test, and it was the first test for every subject (after completing the lineup memory test for the thief, the subjects were subsequently tested for their ability to identify the waiter from a different 8-member simultaneous lineup). Other aspects of the experimental design make it less than ideal for model-testing purposes because the reported data were collapsed across multiple between-subjects experimental conditions (namely high-vs.-low-similarity foils, and biased vs. unbiased instructions). Still, we fit these data for the sake of generality as it is the only other large-*N* study that we know of in which memory for a single target was tested using a lineup. We fit the three models to the data from the first test (involving the thief), and the results are shown in Table 7.

In no case was the fit significantly improved by allowing unequal variance or by the addition of the σ^2_b parameter. All three models are capable of fitting the data without significant deviations, though the Integration model required an extra parameter (σ_{Target} was allowed to differ from 1) to do so. When AIC or BIC is used to judge the relative goodness of fit, the

Ensemble model once again provides the best fit. According to BIC, the Integration model provides the worst fit, whereas according to AIC, the Independent Observations model provides the worst fit.

We also fit these three models to a variety of small-*N* single-target eyewitness identification studies. The results were (perhaps not surprisingly) inconclusive. For example, all three models provided an adequate fit (i.e., non-significant chi-square values) to data reported in Experiment 1 of Mickes, Flowe and Wixted (2012). Similarly, when fit to the fair lineup data from Wetmore et al. (2015), the Ensemble and Independent Observations models adequately fit the data, whereas the Integration model did not. And when fit to the data reported by Carlson et al. (2016), none of the models provided an adequate fit to the data.

General Discussion

We tested three signal detection models that have recently been used to interpret simultaneous lineup performance: the Independent Observations model, the Integration model and the Ensemble model. The Independent Observations model is often used to illustrate how signal detection theory applies to lineups; the Integration model is the most frequently used signal detection model in the eyewitness identification literature to compute *d'*; and the Ensemble model is a quantitative instantiation of a theory of lineup memory recently proposed by Wixted and Mickes (2014). Wixted and Mickes (2014) argued that the simultaneous presentation of similar faces in a lineup facilitates the discounting of non-diagnostic facial features, thereby permitting diagnostic features (those that disproportionately point to the guilty suspect) to play a larger role in the decision. The discounting of non-diagnostic features is an inherent property of the Ensemble model. That is, when memory signals are correlated, subtracting away the ensemble (average) from the MAX face in the lineup removes the contribution of shared (non-diagnostic) features from the diagnostic memory strength variable. The removal of non-diagnostic shared variance reduces error variance and enhances discriminability.[6]

To empirically differentiate between these three models, we first derived their likelihood functions and then fit them to empirical ROC data from multiple eyewitness identification experiments. On balance, the model-fitting evidence would appear to weigh heavily against the Integration model because it generally provided the worst fit. With regard to the other two models, namely, the Independent Observations model and the Ensemble model, both fit the ROC data reasonably well, though a non-trivial edge was apparent for the Ensemble model. The ability of a model to adequately fit the empirical data is important to demonstrate, but, on its own, it does not necessarily validate the best-fitting model (Roberts & Pashler, 2002). Thus, it is also important to consider non-model-fitting evidence bearing on the three models considered here.

Non-Model-Fitting Evidence

As described earlier, the three models make different predictions about the effect of correlated memory signals on discriminability. All three models enjoy the benefit of correlated memory signals on d'_{TP} (the ability to discriminate the guilty suspect from fillers on target-present trials), but they differ in what they predict about d'_{IG} (the ability to discriminate innocent from guilty suspects across trials). The Integration model predicts that correlated memory signals will reduce d'_{IG}; the Ensemble model predicts that correlated memory signals will increase d'_{IG}; and the Independent Observations model predicts no effect of correlated memory signals on d'_{IG}. Because the area under the empirical ROC is jointly determined by both d'_{TP} and d'_{IG}, the Independent Observations and (especially) the Ensemble models predict that the area under the

[6] See Table 1 on page 270 of Wixted & Mickes (2014) for a concrete illustration of this idea in the context of their diagnostic feature-detection theory.

ROC will be higher when decisions are based on correlated memory signals. Several lines of investigation bear on this prediction.

Simultaneous Lineups vs. Sequential Lineups and Showups. A considerable body of recent evidence suggests that the area under the ROC does in fact increase when eyewitness identification procedures allow the eyewitness to take advantage of correlated memory signals. For example, multiple studies have documented a simultaneous lineup advantage over eyewitness identification procedures that present faces in isolation, such as sequential lineups, where the faces in the lineup are presented individually (Carlson & Carlson, 2014; Dobolyi & Dodson, 2013; Gronlund et al., 2012; Mickes et al., 2012), and showups, where only a single face – the suspect – is shown (Gronlund et al., 2012; Mickes, 2015; Wetmore et al., 2015). Because the faces in a simultaneous lineup are likely to be associated with correlated memory signals (in contrast to faces presented in isolation), these results suggest that discriminative performance benefits from correlated memory signals.

Fair vs. Unfair Lineups. The same point applies to the use of fair vs. unfair lineups. In an unfair lineup, the memory signals are less correlated than they otherwise would be because, by definition, the fillers in an unfair lineup do not share features of the perpetrator (only the suspect does). A positive correlation across lineups usually occurs precisely because the features of everyone in the lineup match the features of the perpetrator. Thus, changing that state of affairs, which unfair lineups do, would reduce the correlation. Empirically, unfair lineups have been found to impair the ability to discriminate innocent from guilty suspects (Colloff et al., 2016; Colloff et al., in press), again suggesting that correlated memory signals enhance discriminability.

The fact that performance benefits from correlated memory signals is consistent with both the Independent Observations model and the Ensemble model but weighs against the Integration model, which does not unambiguously predict an advantage of correlated signals and is the only model that is capable of predicting a *disadvantage* under those conditions. A priori, it seems like an odd strategy for eyewitnesses to use (i.e., its prior odds seem low). Quantitatively, it generally does not adequately fit the data; and, qualitatively, its predictions are generally incorrect. Thus, one over-arching conclusion of our investigation is that, going forward, the currently dominant Integration model of lineup memory should probably be abandoned.

Confidence as a Difference Score. Although both the Independent Observations and Ensemble models predict a benefit of correlated memory signals on discriminative performance, they make contrasting predictions about another issue, namely, the degree to which confidence in an ID is affected by the other members of the lineup. In the Independent Observations model, confidence is determined by the memory signal associated with a given face without regard for the other faces in the lineup. In the Ensemble model, confidence is instead determined by the *difference* in the memory signal generated by a face and the ensemble average memory signal of the faces in the lineup. As we noted earlier, studies from a variety of domains have investigated the effect of adding implausible (i.e., dud) alternatives to a set of items on confidence in decisions about the plausible (i.e., non-dud) alternatives in the set. These studies were uniformly interpreted to mean that confidence in plausible alternatives is determined by a difference score (Charman et al. 2011; Hanczakowski et al., 2014; Horry & Brewer, 2016; Windschitl & Chambers, 2004). Findings like these also weigh in favor of the Ensemble model, independent of the model-fitting results reported here.

On the other hand, as we also noted earlier, these results could be explained by assuming that when duds are added to the set, a more liberal decision criterion is used to express high confidence (Hanczakowski et al., 2014). If so, a model that assumes a difference variable (the Ensemble model) would not necessarily fit ROC data better than competing models. The fact that the Ensemble model fared better than the other models in fitting ROC data lends credence to the standard interpretation of the dud effect. That is, the operative memory-strength variable is the degree to which an item in a set of items stands out from the crowd.

Future Directions

The signal detection models considered here do not provide a theoretical account of reaction times. A natural candidate for providing such an account is the Two-stage Dynamic Signal Detection (2DSD) theory proposed by Pleskac and Busemeyer (2010). That model relies on a drift diffusion process to account for choice and decision time at step 1 of the decision-making process and a standard signal detection model to account for confidence at step 2 of the decision-making process. With regard to lineups, decision-making at step 1 would involve detecting the MAX face in the lineup (which is a detection decision that is not usually but could be made explicit). To estimate confidence at step 2, the model assumes that evidence continues to accumulate after the decision at step 1. Confidence at step 2 is theoretically based on this additionally accumulated evidence, which is conceptualized in terms of a standard signal detection model (and which, as here, could be implemented with Independent Observations, Ensemble, or Integration decision variable).

The 2DSD model has not yet been applied to lineup memory, but there is no reason why it could not be. In fact, the main message of our article is that there is no reason why the sophisticated modeling efforts that have been applied to list-memory paradigms should not also be brought to bear on the kind of applied paradigms used in the field of eyewitness identification. In our view, and in the view of others (e.g., Clark, 2003; Clark et al., 2011), basic and applied memory researchers have become far too estranged from each other. The eyewitness identification issues considered here seem too important for that seemingly unnecessary division to remain in place.

References

Albrecht, A. R., & Scholl, B. J. (2010). Perceptually averaging in a continuous visual world: Extracting statistical summary representations over time. *Psychological Science, 21*, 560–567.

Ariely, D. (2001). Seeing sets: Representation by statistical properties. *Psychological Science, 12*, 157–162.

Benjamin, A. S., Diaz, M. L., & Wee, S. (2009). Signal detection with criterion noise: Applications to recognition memory. *Psychological Review, 116,* 84-114.

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

Cameron, E. L., Tai, J. C., Eckstein, M. P., & Carrasco, M. (2004). Signal detection theory applied to three visual search tasks - Identification, yes/no detection and localization. *Spatial Vision*, *17*, 295–325.

Carlson, C. A. & Carlson, M. A. (2014). An evaluation of perpetrator distinctiveness, weapon presence, and lineup presentation using ROC analysis. *Journal of Applied Research in Memory and Cognition*, *3,* 45–53.

Carlson, C. A., Dias, J. L., Weatherford, D. R. & Carlson, M. A. (2016). An investigation of the weapon focus effect and the confidence-accuracy relationship for eyewitness identification. *Journal of Applied Research in Memory and Cognition.* http://dx.doi.org/10.1016/j.jarmac.2016.04.001

Charman, S. D., Wells, G. L. & Joy, S. W. (2011). The dud effect: Adding highly dissimilar fillers increases confidence in lineup identifications. *Law and Human Behavior, 25*, 479-500.

Chong, S. C., & Treisman, A. (2003). Representation of statistical properties. *Vision Research, 43,* 393–404.

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

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

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

Clark, S. E., Moreland, M. B. & Gronlund, S. D. (2014). Evolution of the empirical and theoretical foundations of eyewitness identification reform. *Psychonomic Bulletin & Review, 21*, 251-67.

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

Colloff, M. F., Wade, K. A., Strange, D. & Wixted, J. T. (in press). 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 (2017). *Psychological Science.*

Criss, A. H. & McClelland, J. L. (2006). Differentiating the differentiation models: A comparison of the retrieving effectively from memory model (REM) and the subjective likelihood model (SLiM). *Journal of Memory and Language, 55,* 447–460.

de Fockert, J., & Wolfenstein, C. (2009). Rapid extraction of mean identity from sets of faces. *Quarterly Journal of Experimental Psychology*, *62*, 1716–1722.

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

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

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

Glanzer, M. & Adams, J. K. (1985). The mirror effect in recognition memory. *Memory & Cognition, 13,* 8-20.

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.

Hall, J. F. (1979).Recognition as a function of word frequency. *American Journal of Psychology, 92*, 497-505.

Hanczakowski, M., Zawadzka, K. & Higham, P. (2014). The dud-alternative effect in memory for associations: putting confidence into local context. *Psychonomic Bulletin & Review 21*, 543-548.

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in multiple-trace memory model. *Psychological Review, 95,* 528–551.

Hintzman, D. L. (2001). Similarity, global matching, and judgments of frequency. *Memory & Cognition, 29,* 547–556.

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

Horry, R. & Brewer, N. (2016). How target–lure similarity shapes confidence judgments in multiple-alternative decision tasks. *Journal of Experimental Psychology: General, 145*, 1615-1634.

Lampinen, J. M. (2016). ROC analyses in eyewitness identification research. *Journal of Applied Research in Memory and Cognition, 5,* 21-33.

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. (2002). Signal detection theory. In H. Pashler & J. Wixted (Eds.), *Stevens' handbook of experimental psychology: Methodology in experimental psychology* (3rd ed., pp. 43-90). Hoboken, NJ, US: John Wiley & Sons Inc.

Macmillan N. A. & Creelman, C. D. (1991). *Detection theory: A user's guide*. Mahwah, NJ: Erlbaum.

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

McClelland, J. L., & Chappell, M. (1998). Familiarity breeds differentiation: a subjective-likelihood approach to the effects of experience in recognition memory. *Psychological Review, 105*, 734–760.

Mickes, L. (2015). Receiver operating characteristic analysis and confidence-accuracy characteristic analysis in investigations of system variables and estimator variables that affect eyewitness memory. *Journal of Applied Research in Memory & Cognition , 4*, 93–102.

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

Mickes, L., Seale-Carlisle, T. 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 vs. Confidence Ratings. *Applied Cognitive Psychology, 31,* 467-477.

Neumann, M. F., Schweinberger, S. R., & Burton, A. M. (2013). Viewers extract mean and individual identity from sets of famous faces. *Cognition, 128,* 56–63.

Neuschatz, J. S., Wetmore, S., Key, K. N., Cash, D., Gronlund, S. D. & Goodsell, C. A. (2016). A comprehensive evaluation of showups (pp. 43-69). In Bornstein, B. & Miller, K. (eds.) *Advances in Psychology and Law*. Springer.

Nosofsky, R. M. (1992). Similarity scaling and cognitive process models. *Annual Review of Psychology, 43*, 25-53.

Palmer, E. M., Fencsik, D. E., Flusberg, S. J., Horowitz, T. S., & Wolfe, J. M. (2011). Signal detection evidence for limited capacity in visual search. *Attention, Perception, & Psychophysics, 73*, 2413–2424.

Palmer, J., Verghese, P. & Pavel, M. (2000). The psychophysics of visual search. *Vision Research, 40*, 1227–1268

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

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

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

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

Pleskac T. J. & Busemeyer J. R. (2010). Two-stage dynamic signal detection: a theory of choice, decision time, and confidence. *Psychological Review, 117*, 864–901.

Police Executive Research Forum (2013). A National Survey of Eyewitness Identification Procedures in Law Enforcement Agencies. http://policeforum.org/library/eyewitness-identification/NIJEyewitnessReport.pdf

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

Seale-Carlisle, T. M. & Mickes, L. (2016). US lineups outperform UK lineups. *Royal Society Open Science*. DOI: 10.1098/rsos.160300

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

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

Shrout, P.E. & Fleiss, J.L. (1979). Intraclass Correlations: Uses in Assessing Rater Reliability. *Psychological Bulletin, 2*, 420-428.

Smith, P. L. & Sewell, D. K. (2013). A competitive interaction theory of attentional selection and decision making in brief, multielement displays. *Psychological Review, 120,* 589-627.

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. (in press). Increasing the Similarity of Lineup Fillers to the Suspect Improves the Applied Value of Lineups Without Improving Memory Performance: Commentary on Colloff, Wade, & Strange (2016). *Psychological Science*.

Tulving, E. (1981). Similarity relations in recognition. *Journal of Verbal Learning & Verbal Behavior, 20*, 479-496.

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

Verghese, P. (2001). Visual search and attention: A signal detection theory approach. *Neuron, 31,* 523-535.

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

Windschitl, P. D., & Chambers, J. R. (2004). The dud-alternative effect in likelihood judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*, 198–215.

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-detection model of eyewitness identification. *Psychological Review, 121*, 262-276.

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

Wixted, J. T., Mickes, L., Dunn, J. C., Clark, S. E. & Wells, W. (2016). Estimating the reliability of eyewitness identifications from police lineups. *Proceedings of the National Academy of Sciences, 113*, 304-309.

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

Zawadzka, K., Higham, P. A. and Hanczakowski, M. (2016). Confidence in forced-choice recognition: What underlies the ratings? *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Table 1. Summary of the decision variable and decision rule for each of the three models under consideration here.

Model	**Decision Variable**	**Decision Rule**
Independent Observations	The raw (untransformed) memory strength of a face in the lineup	Identify the MAX face if its memory strength exceeds the decision criterion
Integration	The sum of the memory-strength values across all faces in the lineup	Identify the MAX face if summed memory strength exceeds the decision criterion
Ensemble	The difference between the memory strength of a face and the mean memory strength of the faces in the lineup	Identify the MAX face if its difference score exceeds the decision criterion

Table 2. Results of model-recovery simulations. Panel **A** corresponds to the fits to the uncorrelated data summarized in Figure 9A, and Panel **B** corresponds to the fits to the correlated data summarized in Figure 9B. The parameters μ_{Target}, σ_{Target}, and c_1 through c_6 were free to vary for all fits to determine if their programmed values would be recovered. In addition, σ^2_b was included as a free parameter for the fits to the uncorrelated data if it significantly improved the fit, and, except for the Ensemble model, it was included as a free parameter for the fits to the correlated data (to determine if its programmed value would be recovered). Note that in every case, the model that generated the simulated data, when fit those data, not only fit better than the alternative models but also returned the true programmed parameters (highlighted in bold) with a high degree of accuracy.

A

	Ind Obs Simulated Data				Ensemble Simulated Data				Integration Simulated Data			
Parameter	true	**Ind Obs fit**	ENS fit	INT fit	true	Ind Obs fit	**ENS fit**	INT fit	true	Ind Obs fit	ENS fit	**INT fit**
μ_{Target}	*1.5*	**1.51**	1.47	1.84	*1.5*	0.94	**1.50**	1.75	*1.5*	1.32	1.16	**1.51**
σ_{Target}	*1.0*	**1.01**	1.29	1.68	*1.0*	1.25	**0.97**	2.29	*1.0*	0.80	1.02	**1.09**
c1	*1.2*	**1.20**	1.13	-0.31	*1.1*	0.45	**1.11**	-0.79	*0.0*	1.25	1.15	**-0.04**
c2	*1.4*	**1.40**	1.28	0.46	*1.2*	0.65	**1.21**	-0.06	*2.0*	1.77	1.57	**2.02**
c3	*1.6*	**1.61**	1.44	1.26	*1.5*	1.18	**1.51**	1.84	*3.0*	2.04	1.81	**3.03**
c4	*2.0*	**2.01**	1.79	2.72	*1.8*	1.64	**1.79**	3.44	*4.0*	2.32	2.07	**4.04**
c5	*2.4*	**2.40**	2.16	4.05	*2*	1.94	**1.99**	4.51	*5.0*	2.62	2.37	**5.08**
c6	*2.8*	**2.80**	2.57	5.30	*2.2*	2.22	**2.19**	5.48	*7.0*	3.20	3.00	**7.03**
σ^2_b	*0*	**--**	--	--	*0*	0.73	--	0.14	*0.0*	--	--	**--**
χ^2		**9.5**	814.3	181.3		16.2	**1.47**	40.9		164.1	1499.3	**19.2**
df		**10**	**10**	**10**		**9**	**10**	**9**		**10**	**10**	**10**
p		**0.483**	0.000	0.000		0.064	**0.999**	0.000		0.000	0.000	**0.037**

B

	Ind Obs Simulated Data				Ensemble Simulated Data				Integration Simulated Data			
Parameter	true	**Ind Obs fit**	ENS fit	INT fit	true	Ind Obs fit	**ENS fit**	INT fit	true	Ind Obs fit	ENS fit	**INT fit**
μ_{Target}	*1.5*	**1.50**	1.31	2.61	*1.5*	1.73	**1.49**	3.44	*1.5*	0.85	1.07	**1.50**
σ_{Target}	*1.0*	**1.01**	0.91	1.96	*1.0*	1.18	**0.99**	2.75	*1.0*	0.93	0.82	**0.99**
c1	*1.2*	**1.20**	0.98	1.14	*1.1*	1.40	**1.10**	2.01	*0.0*	0.59	0.88	**-0.01**
c2	*1.4*	**1.39**	1.07	1.88	*1.2*	1.60	**1.20**	2.76	*2.0*	1.00	1.04	**1.99**
c3	*1.6*	**1.59**	1.16	2.62	*1.5*	2.19	**1.49**	4.89	*3.0*	1.20	1.12	**2.99**
c4	*2.0*	**2.01**	1.36	4.20	*1.8*	2.75	**1.80**	6.89	*4.0*	1.40	1.20	**3.95**
c5	*2.4*	**2.42**	1.58	5.70	*2*	3.11	**1.99**	8.16	*5.0*	1.60	1.29	**4.91**
c6	*2.8*	**2.82**	1.80	7.14	*2.2*	3.48	**2.20**	9.45	*7.0*	2.01	1.48	**6.92**
σ^2_b	*0.50*	**0.51**	--	0.13	*0.50*	0.56	--	0.12	*0.50*	0.78	--	**0.48**
χ^2		**8.5**	40.8	11.8		29.4	**5.4**	46.8		7.1	320.2	**2.7**
df		**9**	10	9		9	**10**	9		9	10	**9**
p		**0.485**	0.000	0.225		0.001	**0.867**	0.000		0.631	0.000	**0.976**

Note. “--” means that the parameter was not included in the fit.

Table 3. Results of model-recovery simulations when independent criterion variance was added to lockstep criterion variance. The true values for the criterion parameters now represent the obtained average criterion placement over all simulation trials. The parameters μ_{Target}, σ_{Target}, and c_1 through c_6 were free to vary for all fits. In addition, except for the Ensemble model, σ^2_b was included as a free parameter. For the Ensemble model, its criterion-variance parameter (σ_c) was free to vary as well.

	Ind Obs Simulated Data				Ensemble Simulated Data				Integration Simulated Data			
Parameter	true	**Ind Obs fit**	ENS fit	INT fit	true	BEST fit	**ENS fit**	INT fit	true	BEST fit	ENS fit	**INT fit**
μ_{Target}	*1.5*	**1.40**	1.34	2.52	*1.5*	1.41	**1.52**	2.94	*1.5*	0.83	1.22	**1.48**
σ_{Target}	*1.0*	**1.00**	0.88	1.81	*1.0*	1.26	**1.03**	2.75	*1.0*	0.92	0.82	**0.97**
c1	*0.39*	**0.37**	0.66	-1.79	*0.79*	0.51	**0.80**	-0.59	*0.0*	0.57	0.88	**-0.03**
c2	*0.98*	**0.92**	0.91	0.31	*1.19*	1.25	**1.20**	2.30	*2.0*	0.97	1.20	**1.96**
c3	*1.50*	**1.41**	1.15	2.21	*1.49*	1.83	**1.53**	4.50	*3.0*	1.18	1.36	**2.97**
c4	*2.03*	**1.90**	1.40	4.06	*1.78*	2.34	**1.84**	6.45	*4.0*	1.39	1.52	**3.99**
c5	*2.61*	**2.45**	1.70	6.15	*2.08*	2.85	**2.14**	8.34	*5.0*	1.59	1.68	**5.01**
c6	*3.28*	**3.04**	2.03	8.37	*2.48*	3.48	**2.52**	10.59	*7.0*	1.99	2.00	**6.99**
σ^2_b	*0.50*	**0.56**	--	0.16	*0.50*	0.74	**--**	0.21	*0.50*	0.80	--	**0.50**
σ_c	*0.00*	**--**	0.21	--	*0.00*	--	**0.34**	--	*0.00*	--	0.64	
χ^2		**11.1**	263.8	13.1		62.3	**24.8**	46.4		19.6	197.7	**3.6**
df		**9**	9	9		9	**9**	9		9	9	**9**
p		**0.268**	0.000	0.159		0.000	**0.003**	0.000		0.020	0.000	**0.933**

Table 4. Model fits to confidence-based ROC data from Mickes et al. (2017).

Parameter	**Ind Obs**	**Ensemble**	**Integration**
μ_{Target}	2.01	2.24	3.24
σ_{Target}	0.88	0.67	1.63
c1	1.32	1.39	0.85
c2	1.42	1.46	1.19
c3	1.54	1.54	1.57
c4	1.72	1.67	2.17
c5	1.90	1.80	2.75
c6	2.19	2.02	3.67
c7	2.59	2.32	4.96
c8	3.07	2.69	6.51
σ^2_b	0.32	--	--
χ^2	21.1	**13.4**	27.1
df	13	14	14
p	0.071	0.495	0.019
Ln(L)	-1795.0	-1791.4	-1797.2
AIC	3612.0	**3602.8**	3614.5
BIC	3665.8	**3651.6**	3663.3

Note. *Ln(L)* represents the maximized log likelihood. The lowest χ^2, AIC and BIC values are shown in bold

Table 5. Model fits to instruction-based ROC data from Mickes et al. (2017).

Parameter	**Ind Obs**	**Ensemble**	**Integration**
$\mu_{Target1}$	1.88	2.06	3.04
$\mu_{Target2}$	2.06	2.22	3.44
σ_{Target}	0.95	0.60	1.83
$c1$	0.80	1.08	-0.73
$c2$	1.25	1.36	0.71
$c3$	1.32	1.40	0.91
$c4$	1.63	1.60	1.94
σ^2_b	0.34	--	--
χ^2	10.0	**6.3**	12.5
df	4	5	5
p	0.041	0.278	0.028
Ln(L)	-2852.6	-2850.5	-2853.8
AIC	5721.1	**5715.0**	5723.6
BIC	5771.8	**5759.3**	5766.0

Table 6. Model fits to confidence-based ROC data from Seale-Carlisle & Mickes (2016).

	Ind Obs		**Ensemble**		**Integration**	
Parameter	**US**	**UK**	**US**	**UK**	**US**	**UK**
μ_{Target}	0.97	0.40	1.27	0.60	1.04	0.00
σ_{Target}	1.00	1.00	1.00	1.00	1.78	1.52
c1	0.93	0.90	1.30	1.14	-0.56	-1.40
c2	1.71	1.51	1.83	1.59	2.76	0.98
c3	2.47	2.15	2.45	2.12	5.86	3.32
σ^2_b	0.33	0.06	--	--	0.40	--
χ^2	6.4	10.7	**5.3**	**4.3**	7.7	4.0
df	4	4	5	5	3	4
p	0.172	0.030	0.384	0.506	0.052	0.409
Ln(L)	-1657.7	-1608.7	-1656.9	-1605.3	-1658.3	-1605.1
AIC	3325.3	3227.3	**3321.8**	**3218.7**	3328.5	3220.3
BIC	3350.5	3252.1	**3341.9**	**3238.5**	3358.8	3245.1

Table 7. Model fits to Thief condition of Brewer and Wells (2006).

Parameter	**Ind Obs**	**Ensemble**	**Integration**
μ_{Target}	1.33	1.51	1.30
σ_{Target}	1.00	1.00	2.19
c1	1.56	1.58	0.99
c2	1.61	1.62	1.20
c3	1.75	1.73	1.74
c4	2.07	2.01	2.97
c5	2.71	2.58	5.26
χ^2	12.9	9.0	**8.4**
df	9	9	8
p	0.168	0.436	0.397
Ln(L)	-1747.6	-1745.2	-1745.0
AIC	3507.3	**3502.5**	3504.0
BIC	3537.8	**3533.0**	3539.6

Figure 1. Equal-variance Gaussian signal detection model for lineups. An ID is made if the memory-match signal of the most familiar (MAX) face in the lineup exceeds c_1. In that case, the confidence rating associated with the ID depends on the highest confidence criterion that is exceeded (e.g., the confidence rating is 5 if the strength of the MAX face exceeds c_5). Note that this model corresponds to a fair lineup. In an unfair lineup, the suspect stands out from the other fillers in such a way that the innocent suspect in a target-absent lineup more closely resembles the perpetrator than any of the fillers do. In that case, the innocent suspect and filler distributions would not have the same mean.

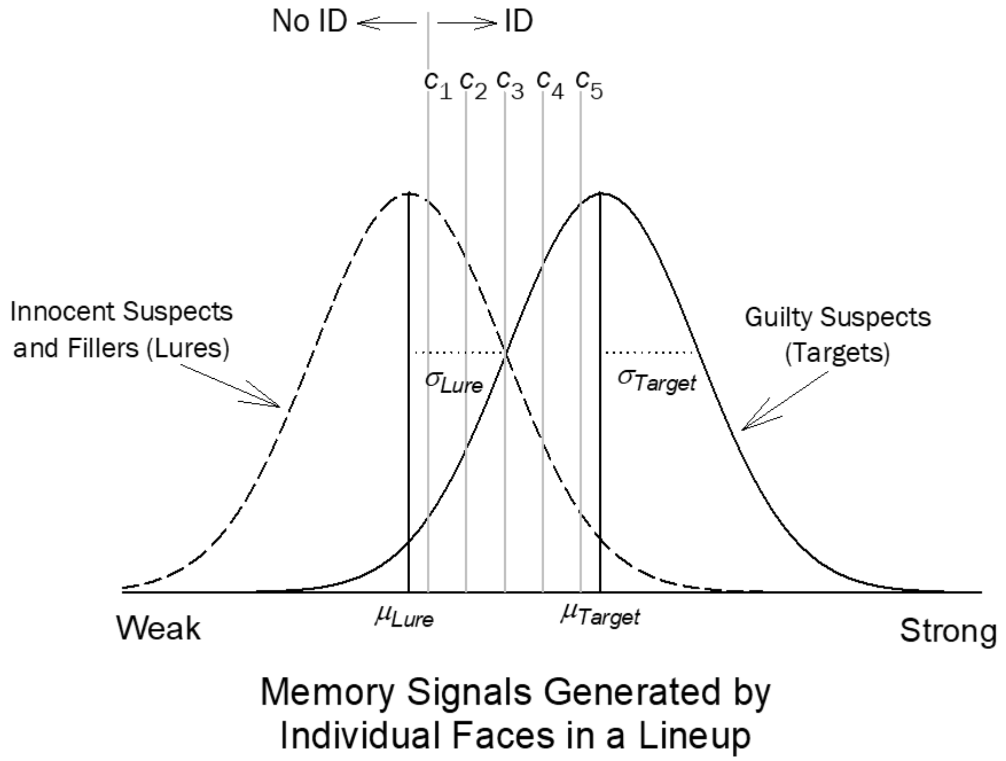

No ID
ID
c_1 c_2 c_3 c_4 c_5
Innocent Suspects and Fillers (Lures)
Guilty Suspects (Targets)
σ_{Lure}
σ_{Target}
Weak
μ_{Lure}
μ_{Target}
Strong
Memory Signals Generated by Individual Faces in a Lineup

Figure 2. Simulated distributions of the decision variable under the Independent Observations model (i.e., raw memory signals) for innocent and guilty suspects as a function of the correlation of memory signals between suspects and lures (ρ). (A) Shows these distributions across all lineups, (B) shows lineups conditional on the suspect generating the maximum memory signal (i.e., the memory signal for the suspect was greater than that of all the lures). The obtained (simulated) d'_{IG} in panel **A** remains constant at its programmed value of 2.0 as ρ increases. The distributions in panel **B** are frequency distributions, which show that IDs of guilty suspects from target-present lineups increase as the correlation increases. The numbers above each distribution indicate the proportion of target-present and target-absent trials in which the guilty suspect or innocent suspect, respectively, generated the MAX signal in the simulation. These numbers therefore represent the maximum hit and false alarm rates, and they illustrate the fact that, all else being equal, as ρ increases, the ability to discriminate the guilty suspect from the fillers in target-present increases (i.e., d'_{TP} increases), selectively increasing the correct ID rate.

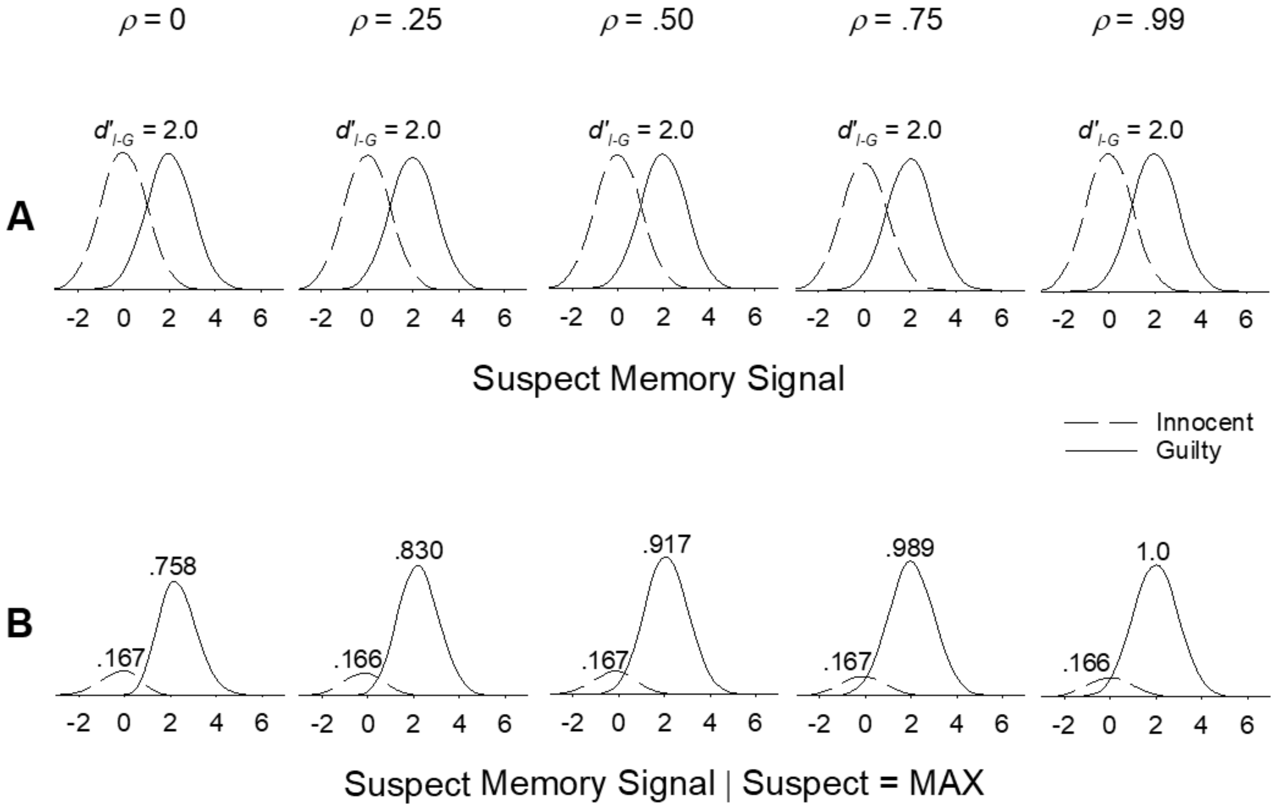


Figure 3. Integration signal detection model for lineups. An ID is made if the summed memory-strength of the faces in the lineup exceeds the decision criterion (with confidence determined by the highest criterion exceeded). When the summed memory strength signal exceeds the criterion, the face that is identified is the face that generates the MAX (non-summed) memory signal in the lineup. Note that μ_{Target} and μ_{Lure} here are the same as μ_{Target} and μ_{Lure} for the untransformed memory signals in Figure 1 (i.e., the summing operation does not change these mean values). However, the standard deviations depicted here (σ_{TP} and σ_{TA}) differ from the standard deviations for the untransformed memory signals (σ_{Target} and σ_{Lure}) in Figure 1 because the variance of a summed (uncorrelated or positively correlated) random variable is greater than the variance of the individual components.

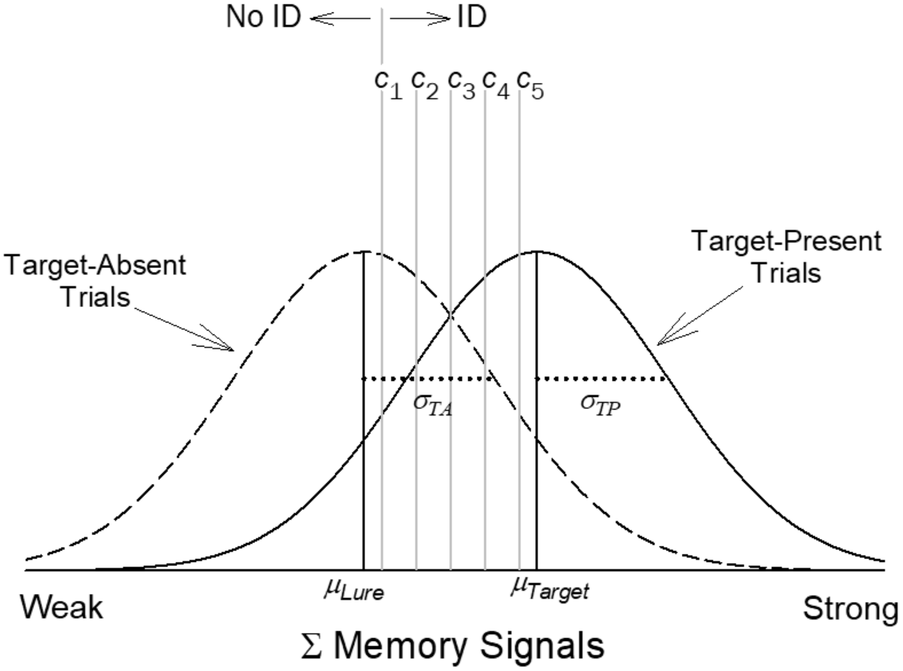


Figure 4. Simulated distributions of the decision variable under the Integration model (i.e., summed memory signals) as a function of the correlation of memory signals between suspects and lures (ρ). (A) Shows these distributions across all lineups, (B) shows the lineups conditional on the suspect generating the maximum memory signal (i.e., the memory signal for the suspect was greater than that of all the lures). The distributions in panel **B** are frequency distributions, and the number labels show the proportion of trials in which the suspect yielded the MAX value. Except for random error, the proportions are the same as those shown in panel **B** of Figure 2, but the distribution of the summed diagnostic decision variable on those trials is different.

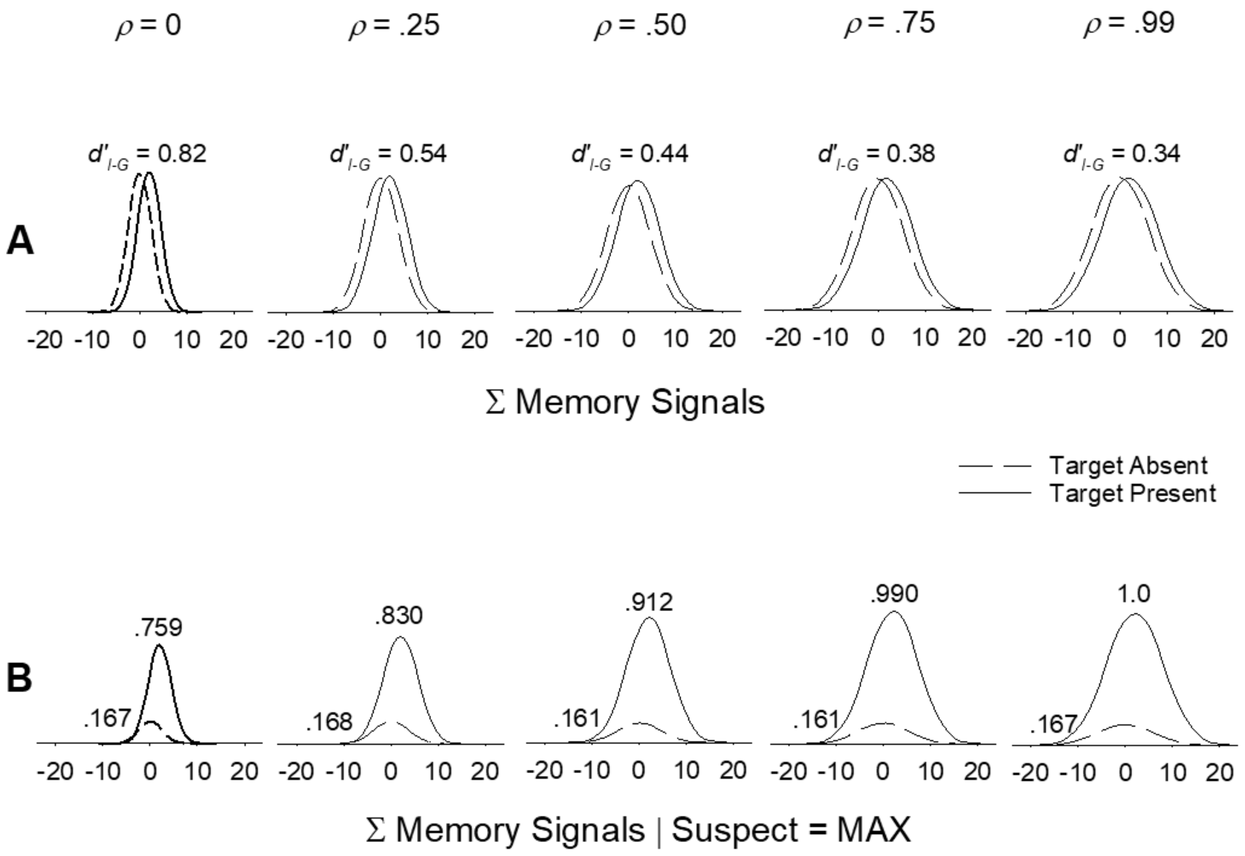


Figure 5. Ensemble signal detection model for lineups of size *k*. An ID is made if the difference between memory signals of a face minus the ensemble average (μ) exceeds the decision criterion (with confidence determined by the highest criterion exceeded). When the difference score exceeds the criterion, the face that is identified is the face that generates the MAX difference score. The standard deviations for the target-absent and target-present lineups here (σ_I and σ_G, respectively) differ from the corresponding standard deviations for the untransformed memory signals in Figure 1 (σ_{Lure} and σ_{Target}, respectively). Here, smaller standard deviations are shown, corresponding to a positive correlation.

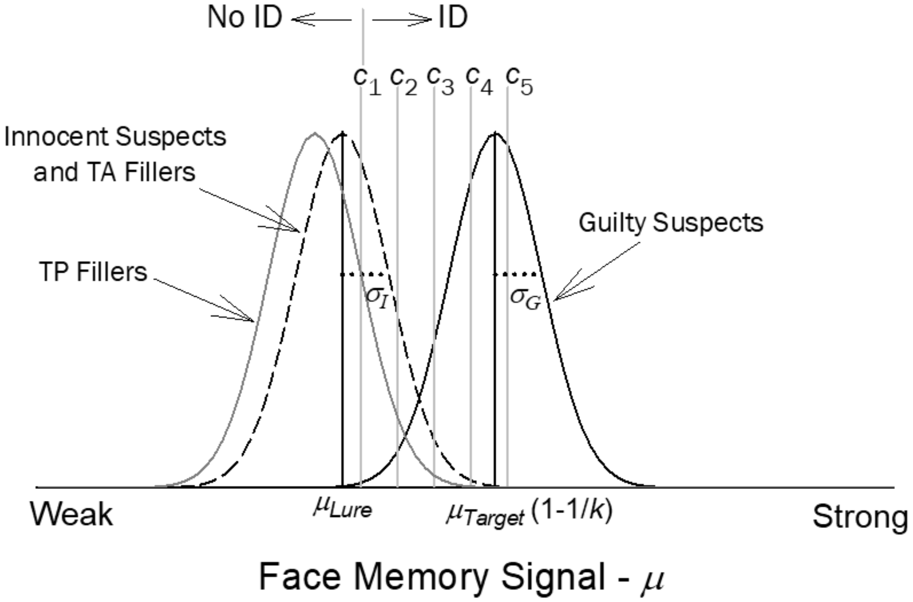


Figure 6. Simulated distributions of the decision variable under the Ensemble model (i.e., raw memory signal minus mean memory signal) for innocent and guilty suspects as a function of the correlation of memory signals between suspects and lures (*ρ*). (A) Shows these distributions across all lineups, (B) shows lineups conditional on the suspect generating the maximum memory signal (i.e., the memory signal for the suspect was greater than that of all the lures). The distributions in panel **B** are frequency distributions, and the number labels show the proportion of trials in which the suspect yielded the MAX value. Except for random error, the proportions are the same as those shown in panel **B** of Figures 2 and 4, but the distribution of the diagnostic decision variable on those trials is different.

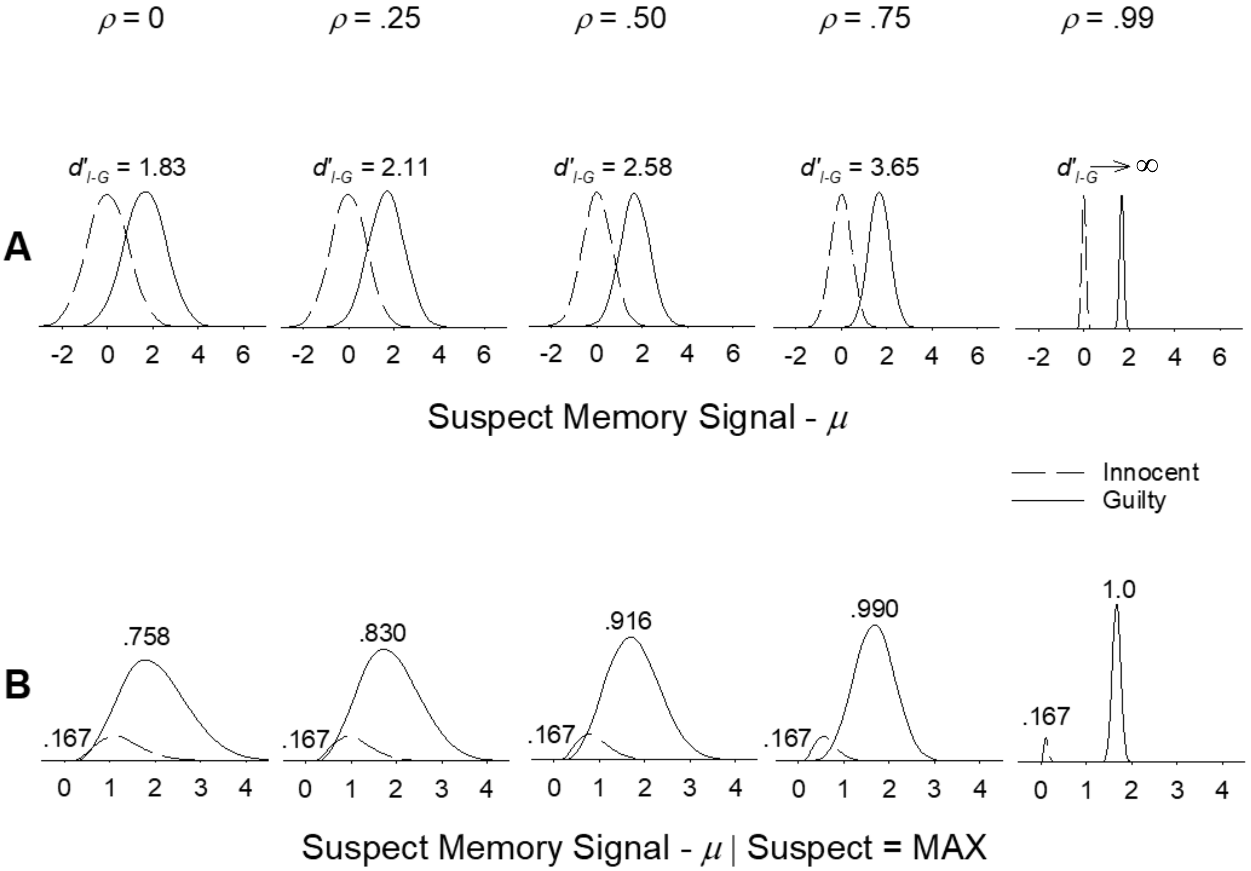


Figure 7. This figure (center plot) shows the joint distribution of suspect (*x*) and filler (*y*) memory strengths for 2-alternative target-present (black), and target-absent (grey) trials. The different decision rules can be thought of as collapsing these joint distributions in different ways: the independent observations decision variable for guilty and innocent suspects just amounts to the distribution along *x* for target-present and target-absent trials. The integration decision variable calculates suspect+filler for each trial, and thus marginalizes along one diagonal of the suspect,filler distribution. The Ensemble model (technically here, BEST minus REST), uses the difference of suspect-lure memory strengths as the decision variable, and thus marginalizes the joint distribution along the other diagonal. In the presence of a correlation (here, $\rho = 0.8$) of suspect and filler signals on a given trial, the Ensemble (difference) variable clearly yields a greater separation between guilty and innocent suspects than the "independent" signal alone; moreover, the "integration" (additive) variable clearly yields lower separation than the independent observations decision variable.

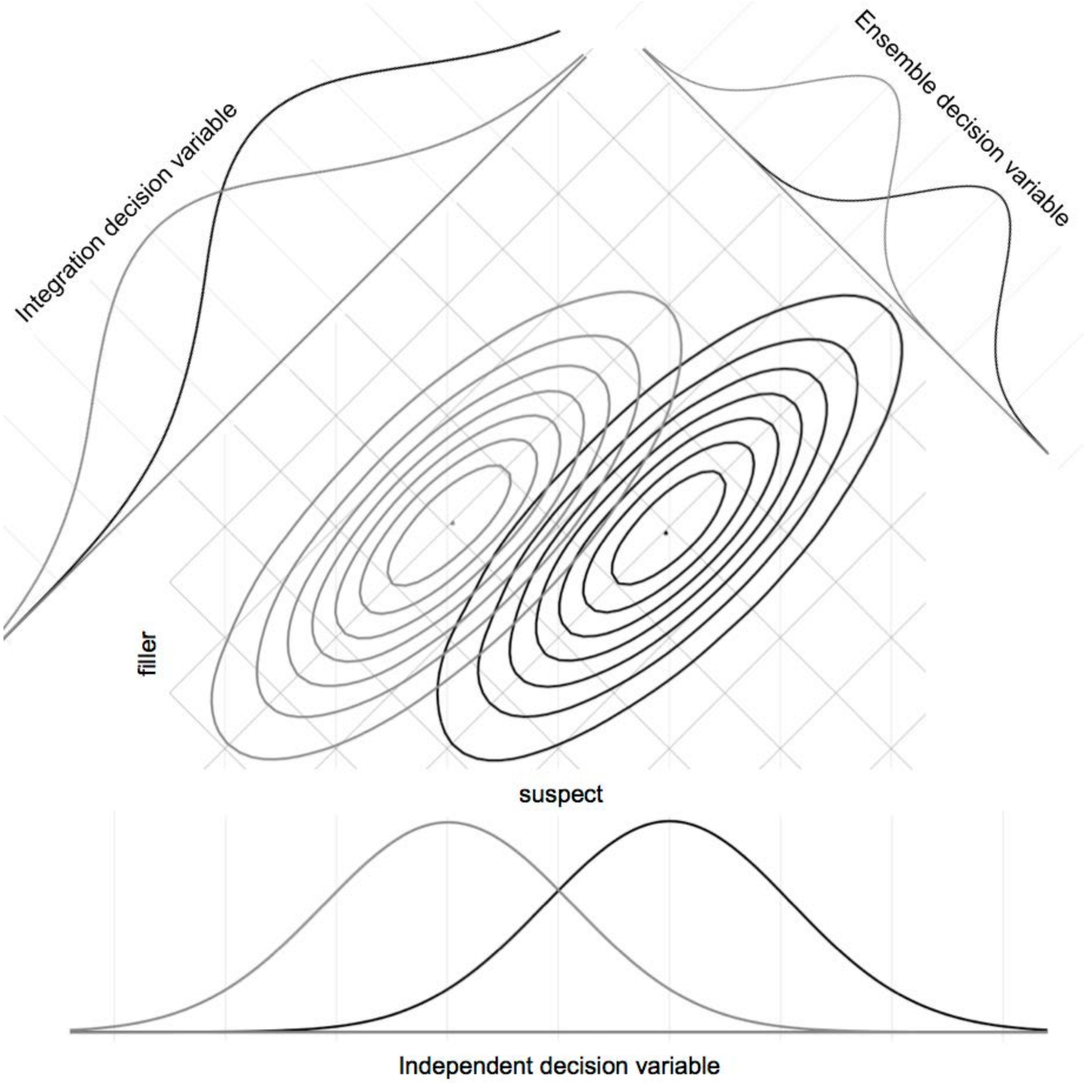

Integration decision variable
Ensemble decision variable
filler
suspect
Independent decision variable

Figure 8. Distribution of memory-match signals across lineups (solid distributions = guilty suspects; dashed distributions = innocent suspects and fillers) as the correlation (*ρ*) increases from 0 to 1. The net distributions shown in the bottom row (Row 5) are all the same and correspond to memory strength distributions aggregated across trials (as in Figure 1). Rows 1 through 4 shows the distributions from which innocent and guilty suspect values are drawn for 4 separate lineups. The columns correspond to an increasing correlation such that more and more of the variance in the aggregate distributions is accounted for by between-lineup variance and less and less by within-lineup variance.

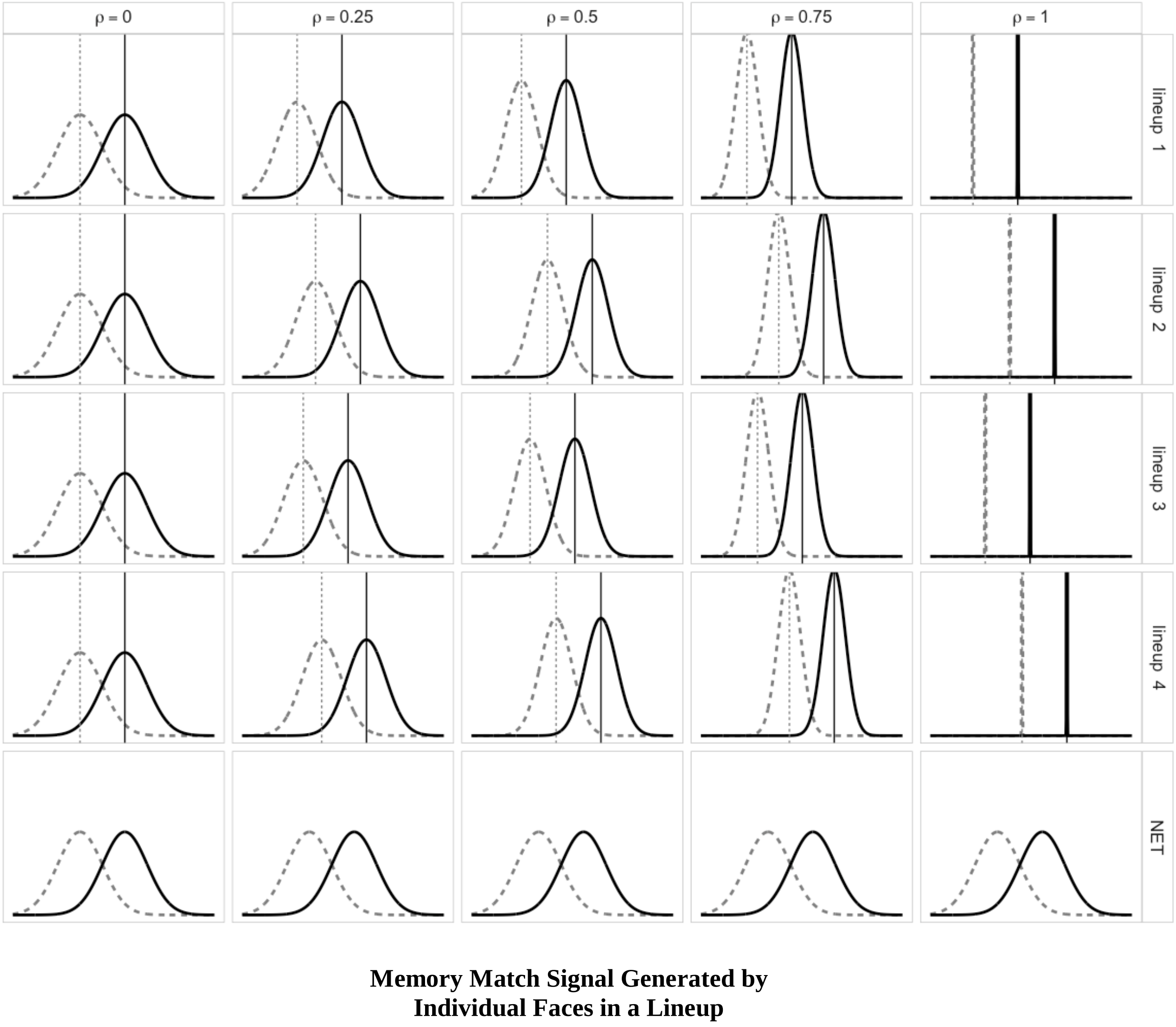


Figure 9. Results from two representative model-recovery simulations in which simulated data were generated by each of the three models (as shown on the x-axis) and then the three models were fit to each data set (the fitted model is shown in the legends). In panel **A**, the correlation in the programmed raw memory signals was set to 0 (i.e., $\sigma^2{}_b = 0$), as in Column 1 of Figure 8. Because $\sigma^2 = \sigma^2{}_b + \sigma^2{}_w$, and because we set $\sigma^2 = 1$ for these equal-variance simulations, this means that $\sigma^2{}_w = 1$. In panel **B**, the correlation in the programmed raw memory signals was set to .5 (i.e., $\sigma^2{}_b = .5$), as in Column 3 of Figure 8, which means that $\sigma^2{}_w = .5$ as well. The models that generated the simulated data (consisting of 10,000 target-present and 10,000 target-absent trials) fit better than the alternative models in every case. The programmed (and estimated) parameters for these fits are presented in Table 2.

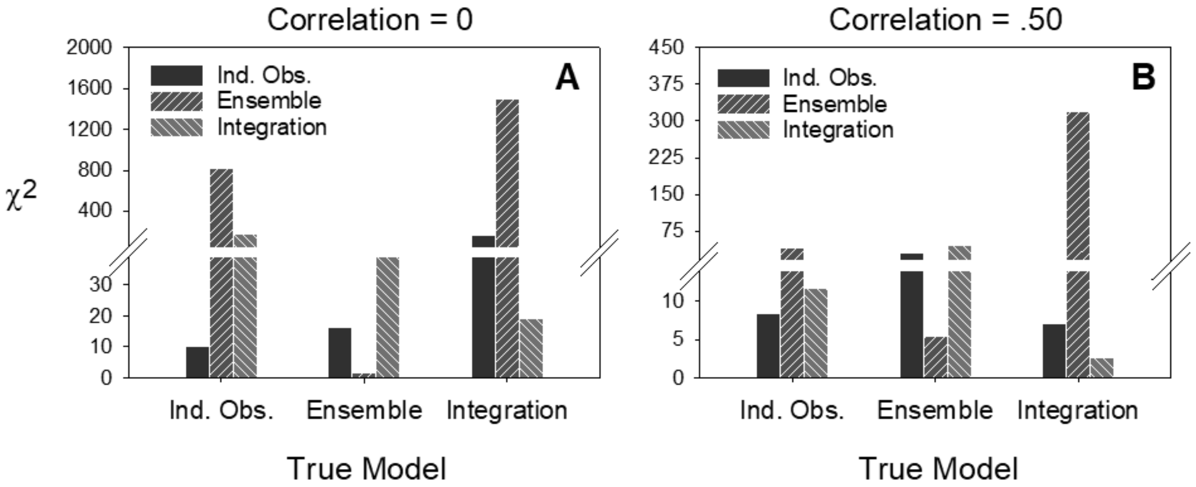


Figure 10. Confidence-based and instruction-based ROC data from Mickes et al. (2017).

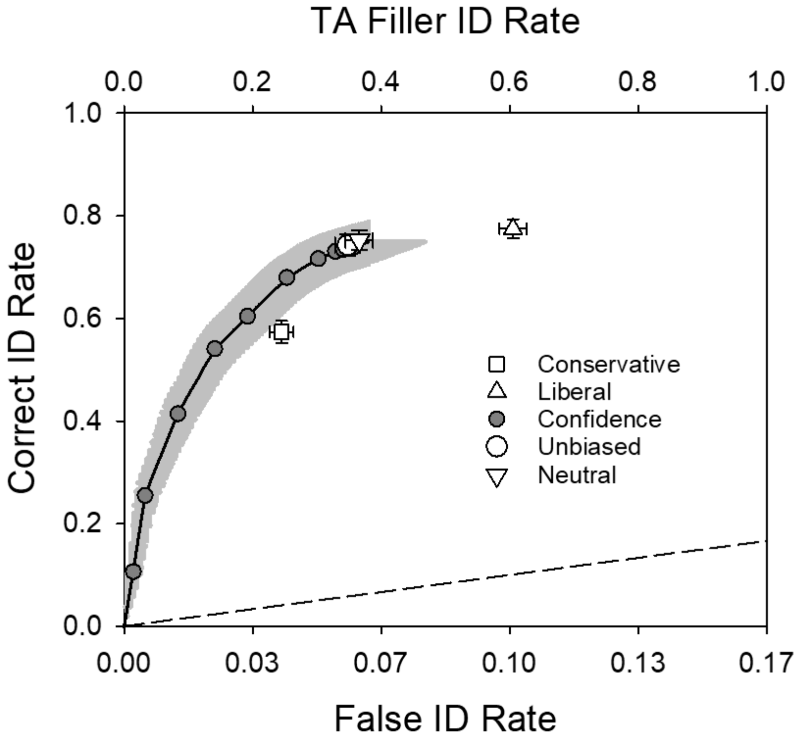


Figure 11. **A**. Observed ROC data from the confidence condition of Mickes et al. (2017) and ROC data predicted by the three competing models using their maximum-likelihood parameter estimates (TP = target-present and TA = target-absent). **B**. Observed and predicted target-present ROC data, with the target-present filler ID rate now plotted on the *x*-axis.

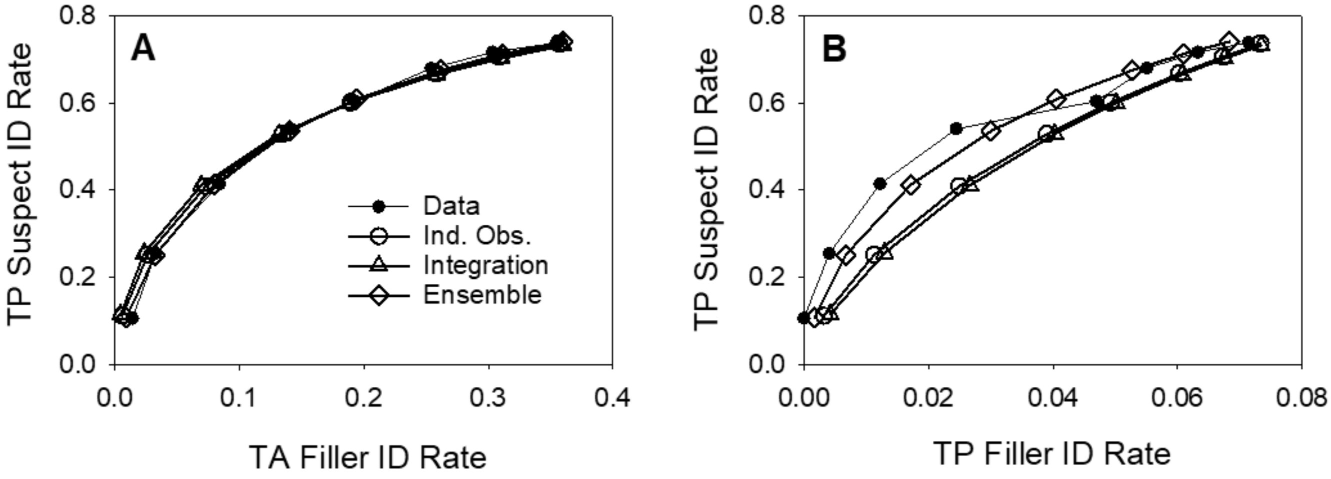
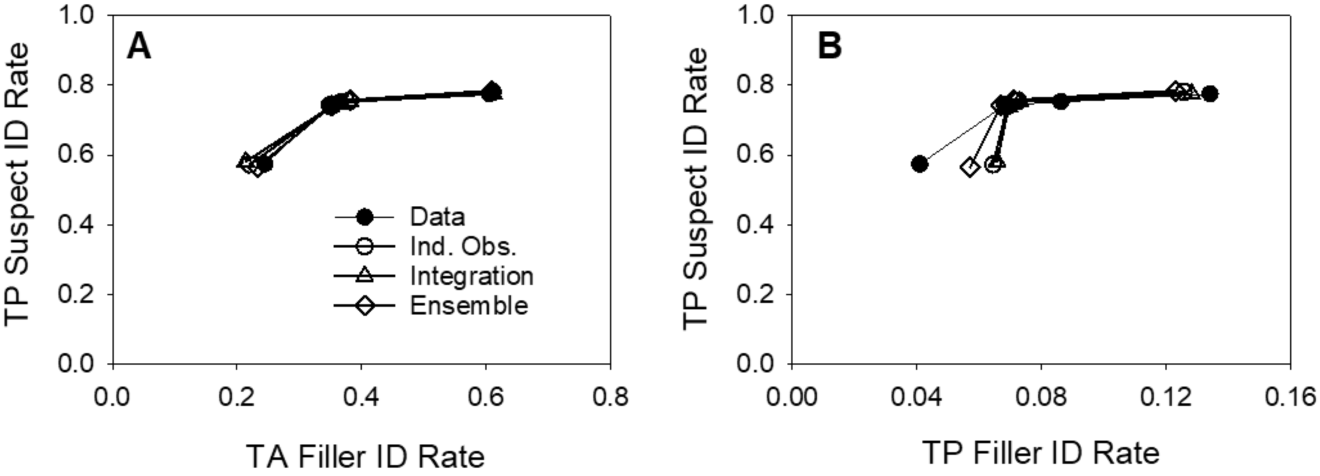


Figure 12. **A**. Observed ROC data from the instructional-biasing condition of Mickes et al. (2017) and ROC data predicted by the three competing models using their maximum-likelihood parameter estimates (TP = target-present and TA = target-absent). **B**. Observed and predicted target-present ROC data, with the target-present filler ID rate now plotted on the *x*-axis.



Appendix A: Integration and Ensemble models for k-alternative lineups

Integration Model

To appreciate how this model works, consider first the simplest case of a 2-alternative lineup. A target-present trial would consist of a guilty suspect value drawn from a distribution with mean and standard deviation of μ_{Target} and σ_{Target}, respectively, and a filler drawn from a distribution with mean and standard deviation of μ_{Lure} and σ_{Lure}, respectively. Thus, setting μ_{Lure} = 0, the mean of the summed distribution on target-present trials would be μ_{Target} + μ_{Lure} = μ_{Target} + 0 = μ_{Target}, and the variance of that summed distribution would be σ^2_{Target} + σ^2_{Lure} + $2\rho\sigma_{Target}$ σ_{Lure}, where ρ represents the correlation between the two memory signals across trials. As described earlier, $\rho = \sigma^2_b$ / ($\sigma^2_b + \sigma^2_w$). Note that the larger the correlation, the greater the variance of the summed decision variable. Similarly, a fair target-absent trial would consist of an innocent suspect drawn from a distribution with mean and standard deviation of μ_{Lure} and σ_{Lure}, respectively, and a filler drawn from the same distribution. Thus, the mean of the summed distribution on target-absent trials would be μ_{Lure} + μ_{Lure} = 0 + 0 = 0, and the variance of that summed distribution would be σ^2_{lure} + σ^2_{lure} + $2\rho\sigma_{Lure}$ σ_{Lure}.

Because the mean of the summed variable on target-present trials is μ_{Target} and the mean of the summed variable on target-absent trials is 0, the numerator of the d' measure for this model is μ_{Target} – 0 = μ_{Target}. For the equal-variance version of the model, σ^2_{Target} = σ^2_{Lure} = σ^2, in which case the common variance of the target-present and target-absent distributions in the denominator would be $2\sigma^2 + 2\rho\sigma^2$. Setting σ^2 = 1, the variance of the two distributions becomes 2 + 2ρ, or 2(1 + ρ). Thus, discriminability between innocent and guilty suspects would be equal to $d' = \mu_{Target}/\sqrt{2(1+\rho)}$. According to this equation, discriminability for the Integration model is lower than it would be for the Independent Observations model (for which d' = μ_{Target}) even when ρ = 0, and it only gets worse as the correlation increases.

The same basic message applies to larger lineups of size k. The mean of the summed random variable on target-present trials is the sum of the means of the components, or μ_{Target} + $\sum_1^{k-1} \mu_{Lure}$ = μ_{Target} + (k-1)μ_{Lure}, where the sum reflects the fact that there are k – 1 fillers in the target-present lineup. On target-absent trials, the mean is simply μ_{Lure} + (k-1)μ_{Lure} = $k\mu_{Lure}$. Because we set μ_{Lure} = 0 by convention, the means of the summed memory-strength variables on target-present and target-absent trials are equal to μ_{Target} and 0, respectively. For the uncorrelated equal-variance case (where σ^2_{Target} = σ^2_{Lure} = σ^2), the sum of the k component variances is simply $k\sigma^2$, and this is true for both target-present and target-absent lineups. For correlated random variables, the variance of the sum, Var(Sum), is given by

$$\text{Var(Sum)} = k\sigma^2 + \sum_{i=1}^{k} \sum_{j \neq i} \rho \sigma_i \sigma_j$$

In the equal-variance version of the model, this equation reduces to

$$V\text{ar(Sum)} = k\sigma^2 + k(k-1)\rho\sigma^2$$

or, after setting σ^2 = 1 and rearranging:

$$V\text{ar(Sum)} = k[1 + (k-1)\rho]$$

Thus, discriminability for a k-alternative lineup according to the INTEGRATION model is:

$$d' = \mu_{Target}/\sqrt{k[1 + (k-1)\rho]}$$

According to this model, as ρ increases, discriminability should decrease. In the uncorrelated (ρ = 0) case, discriminability for the INTEGRATION model becomes:

$$d' = \mu_{Target}/\sqrt{k}$$

Ensemble Model

Let *x* be a random variable from the target distribution, *y* be a random variable from the ensemble average distribution on target-present trials of a k-alternative lineup, and *z* be a random variable from the lure distribution. In that case,

$$\mathrm{Var}(x) = \sigma^2_{Target}$$

$$\mathrm{Var}(y) = \left(\sigma^2_{Target} + (k-1)\sigma^2_{Lure} + \sum_{i=1}^{k}\sum_{j\neq i}\rho\sigma_i\sigma_j\right)\Big/k^2$$

$$\mathrm{Var}(z) = \sigma^2_{Lure}$$

For the equal-variance case, $\sigma^2_{Target} = \sigma^2_{Lure} = \sigma^2$, which simplifies the expression for Var(*y*):

$$\mathrm{Var}(y) = \left[\sigma^2 + \sum \sigma^2 + k(k-1)\rho\sigma^2\right]\Big/k^2$$

$$\mathrm{Var}(y) = [k\sigma^2 + k(k-1)\rho\sigma^2]/k^2$$

$$\mathrm{Var}(y) = \sigma^2[1 + \rho(k-1)]/k$$

To compute the variance of $x – y$ (which is the decision variable according to the Ensemble model), we eventually make use of this definitional formula:

$$\mathrm{Var}(x-y) = Var(x) + Var(y) - 2Cov(x,y)$$

To compute Cov(*x*,*y*) in the formula above, we also make use of this definitional formula:

$$Cov(x,y) = E(xy) - E(x)E(y).$$

The *E(x)* and *E(y)* components of this covariance formula are straightforward and are given by:

$$E(x) = \mu_{Target}$$

$$E(y) = \left(\mu_{Target} + \sum \mu_{Lure}\right)\Big/k$$

where ∑ (here and below) means to sum over k – 1 fillers.

$$E(y) = \mu_{Target}/k$$

E(*xy*) is given by:

$$E(xy) = E\left[x\left(x + \sum z\right)\ / \ k\right] = E\left[\ x^2/k + \sum xz/k\right] = E[x^2/k] + E\left[\sum xz/k\right]$$

The two terms on the right of the above expression equal

$$E[x^2/k] = \ \left[\mu_{Target}^2 + \sigma_{Target}^2\right]/k$$

and

$$E\left[\sum xz/k\right] = \sum\left[Cov(x,z) + \mu_{Target}\mu_{Lure}\right]/k$$

$$E\left[\sum xz/k\right] = Cov(x,z)(k-1)/k$$

Thus,

$$E(xy) = E[x^2/k] + E\left[\sum xz/k\right] = \left[\mu_{Target}^2 + \sigma_{Target}^2\right]/k + Cov(x,z)(k-1)/k$$

For equal-variance case, this expression reduces to:

$$E(xy) = \left[\mu_{Target}^2 + \sigma^2\right]/k + Cov(x,z)(k-1)/k$$

Cov(*x*,*z*) in the above expression is just the covariance between random variables drawn from the target and lure distributions:

$$Cov(x,z) = \rho\sigma_{Target}\sigma_{Lure}$$

In the equal-variance case, this equation becomes:

$$Cov(x,z) = \rho\sigma^2$$

Thus:

$$E(xy) = \left[\mu_{Target}^2 + \sigma^2\right]/k + \rho\sigma^2(k-1)/k$$

$$E(xy) = \left[\mu_{Target}^2 + \sigma^2 + \rho\sigma^2(k-1)/k\right]$$

The values computed above can now be plugged into:

$$Cov(x,y) = E(xy) - E(x)E(y)$$

$$Cov(x,y) = \left[\mu_{Target}^2 + \sigma^2 + \rho\sigma^2(k-1)/k\right] - \mu_{Target}\left(\mu_{Target}/k\right)$$

$$Cov(x,y) = \left[\mu_{Target}^2 + \sigma^2 + \rho\sigma^2(k-1) - \mu_{Target}^2\right]/k$$

$$Cov(x,y) = [\sigma^2 + \rho\sigma^2(k-1)]/k$$

$$Cov(x,y) = \sigma^2[1 + \rho(k-1)]/k$$

Now we are in a position to compute the variable of interest, namely, the variance of the *x* – *y* decision variable:

$$Var(x-y) = Var(x) + Var(y) - 2Cov(x,y)$$

For the equal-variance case,

$$Var(x) = \sigma^2$$

$$Var(y) = \sigma^2[1 + \rho(k-1)]/k$$

$$Cov(x,y) = \sigma^2[1 + \rho(k-1)]/k$$

Thus,

$$Var(x-y) = \sigma^2 + \sigma^2[1 + \rho(k-1)]/k - 2\,\sigma^2[1 + \rho(k-1)]/k$$

$$Var(x-y) = \sigma^2 - \sigma^2\left(1 + \rho(k-1)\right)/k$$

$$Var(x-y) = \sigma^2\left[1 - \left(1 + \rho(k-1)\right)/k\right]$$

$$Var(x-y) = \sigma^2\left(k - 1 + \rho(k-1)\right)/k$$

$$Var(x-y) = \sigma^2(k-1)(1-\rho)/k$$

$$Var(x-y) = \sigma^2(1-\rho)(1-1/k)$$

Setting $\sigma^2 = 1$ yields the final result for the variance of the *x* – *y* decision variable:

$$Var(x-y) = (1-\rho)(1-1/k)$$

This variance estimate can be used to compute *d'* because the square root of that value is the denominator of the *d'* formula. The numerator of the *d'* formula is the difference between the mean of *x*, which is equal to μ_{Target}, and the mean of *y*, which is equal to $(\mu_{Target} + \sum\mu_{Lure}) / k$.

Because μ_{Lure} = 0, the difference in the numerator reduces to μ_{Target} - μ_{Target} / k. This can also be written [(k-1)/k]μ_{Target}, or (1-1/k) μ_{Target} Thus,

$$d' = (1 - 1/k)\mu_{target}/\sqrt{(1 - 1/k)(1 - \rho)}$$

$$d' = \mu_{target}\sqrt{(1 - 1/k)}/\sqrt{(1 - \rho)}$$

$$d' = \mu_{target}/\sqrt{[k/(k - 1)](1 - \rho)}$$

Appendix B: Likelihoods

Table of Contents

Goal

Our goal in this section is to derive the likelihood functions for a memory decision on various trials under different decision models. These will take the form:

$$\mathcal{L}_f(r, c \mid \theta, \mathbf{q})$$

- f indicates the memory/decision model in question (here we will consider different 'decision variable' functions that people might use in an identification task: BEST, BEST-REST, BEST-ENSEMBLE, INTEGRATION).

- r is the response (identified item): either none ($\emptyset$) or the index of the item identified (i), or the class of the item ((T)arget, (L)ure).

- c is the confidence level of that identification.

- θ corresponds to the parameters of the model – the summary statistics of the target and lure memory distributions (which we will expand on below).

- $\mathbf{q}$ is the trial specification: a vector of length k (the number of items present on a trial), with each element q_i indicating whether that item was a target (T) or a lure (L). On a target-absent trial, all elements of $\mathbf{q}$ are L, while in a target present trial, the first element is marked as the target ($q_1 = \mathrm{T}$).

Partitioning correlations into independent sources.

Before we start with the derivations, it is useful to explain the mathematical isomorphism between considering many variables and their (homogenous) pairwise correlations, and factoring that representation into independent and shared sources of variance.

Let b be a random sample of the variability *shared* by all items on a given trial (the between-trial variability), distributed with mean 0 and variance σ_b^2. Let w_i be a random sample of the *independent* variability for item i on that trial j, distributed with mean μ_{w_i} and variance $\sigma_{w_i}^2$. The net memory signal x_i for item i on that trial is the sum of b and w_i, so $x_i = b + w_i$, with mean $\mu_{x_i} = \mu_{w_i}$ and variance $\sigma_{x_i}^2 = \sigma_b^2 + \sigma_{w_i}^2$.

The covariance of the net memory strength for two items on a given trial (x_i and x_j) is determined entirely by their shared variability: $\sigma_{x_i,x_j} = \sigma_b^2$, and their correlation is given by $\rho_{x_i,x_j} = \frac{\sigma_b^2}{\sigma_{x_i}\sigma_{x_j}} = \frac{\sigma_b^2}{\sqrt{\sigma_b^2+\sigma_{w_i}^2}\sqrt{\sigma_b^2+\sigma_{w_j}^2}}$.

Given the isomorphism between the ‘shared variability’ formulation and the ‘marginal correlation’ formulation, it is possible to carry out the subsequent derivations using either (a) the marginal net strengths of different items (x_i), their variances ($\sigma_{x_i}^2$) and pairwise correlations ($\rho_{x_i,x_{i'}}$), or
(b) using the shared (b) and independent (w_i) perturbations in memory strength, and their variances (σ_b^2 and $\sigma_{w_i}^2$).

Since we can derive (a) from (b), and vice versa, they are mathematically indistinguishable, but we find the math to be more concise when pursuing option (b), so we will mostly lay out derivations in terms of those variables. However, where we find it useful, we will translate into marginal variances and pairwise item-item correlations.

Memory decisions on a given trial.

We assume that identifying a given item i at confidence level c requires that
(a) item i has the highest memory signal on that trial, and
(b) the net "decision variable" is above c

The "decision variable" is a function of the memory signals from all k items in that trial ($f(\mathbf{x})$), and what that function is differs based on the model (the "decision rule"):
- for the "BEST" model, $f(\mathbf{x}) = \max(\mathbf{x})$
- for the "BEST-REST" model, $f(\mathbf{x}) = \max(\mathbf{x}) - \frac{1}{k-1}\sum_{x_i \neq \max(\mathbf{x})} x_i$
- for the "BEST-ENSEMBLE" model, $f(\mathbf{x}) = \max(\mathbf{x}) - \frac{1}{k}\sum_i x_i$
- for the "INTEGRATION" model, $f(\mathbf{x}) = \sum_i x_i$

Generic likelihoods

We can write out fairly generic likelihoods for all of these models. The probability that a particular item with memory strength (x_i) is identified as the target at confidence threshold c is given by

$$\mathcal{L}_f(r = i, c \mid \theta, \mathbf{q}) = P(x_i = \max(\mathbf{x}), f(\mathbf{x}) > c)$$

We decompose this using the chain rule and the law of total probability into:

$$P(x_i = \max(\mathbf{x}), f(\mathbf{x}) > c) = \\ \int_{-\infty}^{\infty} P\,(x_i) \\ P(x_i = \max(\mathbf{x}) \mid x_i) \\ P(f(\mathbf{x}) > c \mid x_i, x_i = \max(\mathbf{x})) \quad dx_i$$

The probability that a given x_i is the largest element of $\mathbf{x}$ is equal to the probability that all other elements of $\mathbf{x}$ are smaller than x_i. Since the shared trial variance (w) is a constant offset added to all items ($x_i = w_i + b$), the comparison of x_i to all other items can be carried out by simply considering the independent variability for those items ($\mathbf{w} = \mathbf{x} - b$). Consequently, the probability that a given x_i is the probability that the corresponding w_i is larger than all other w_js. Moreover, because the w_is are, by definition, independent for all elements of $\mathbf{w}$, the probability that $w_j < w_i$ for all j is the product across all js:

$$P(x_i = \max(\mathbf{x}) \mid x_i) = P(w_i = \max(\mathbf{w}) \mid w_i) = \prod_{j \neq i} P\,(w_j < w_i \mid w_i)$$

Furthermore, under the standard signal detection theory formulation, we assume that the variables are all normally distributed with a probability density function of $n(x \mid \mu, \sigma)$, and a cumulative distribution of $N(x \mid \mu, \sigma)$. This allows us to formally write out that:

$$P(w_i) = n(w_i \mid \mu_{w_i}, \sigma_{w_i})$$

Moreover, since $P(w_j < w_i \mid w_i) = N(w_i \mid \mu_{w_j}, \sigma_{w_j})$. We get:

$$P(x_i = \max(\mathbf{x}) \mid x_i) = \prod_{j \neq i} N\left(w_i \mid \mu_{w_j}, \sigma_{w_j}\right)$$

Substituting both of these into our earlier equation, we get the following expression for the probability that a given item i will be identified as the target on a given trial above a certain confidence level:

$$\mathcal{L}_f(r = i, c \mid \theta, \mathbf{q}) = \int_{-\infty}^{\infty} n\,(w_i|\mu_{w_i}, \sigma_{w_i}) \prod_{j \neq i} N\left(w_i \mid \mu_{w_j}, \sigma_{w_j}\right) P(f(\mathbf{x}) > c \mid w_i, w_i = \max(\mathbf{w}))\, dw_i$$

Next, rather than referring to unique μ_{w_i} and σ_{w_i} for all is, we will rely on the fact that these are the same for all items that are the same type (target or lure) and that $\mathbf{q}$ encodes the type of item via each q_i. Thus, we can explicitly rewrite these in terms of μ_T, ($\mu_L = 0$), σ_T, and σ_L by relying on the indicator $\mathbf{q}$ by using the notation $\mu_{q[i]}$ and $\sigma_{q[i]}$ to refer to the mean and standard deviation for lures or targets, (depending on the type of item $q[i]$).

$$\mathcal{L}_f(r = i, c \mid \theta, \mathbf{q}) = \int_{-\infty}^{\infty} n\,(w_i|\mu_{q[i]}, \sigma_{q[i]}) \prod_{j \neq i} N\left(w_i \mid \mu_{q[j]}, \sigma_{q[j]}\right) P(f(\mathbf{x}) > c \mid w_i, w_i = \max(\mathbf{w})) \quad dw_i$$

The third term of the integrand ($P(f(\mathbf{x}) > c \mid w_i, w_i = \max(\mathbf{w}))$) depends on the specific decision-variable model. Below, we show that for all models we consider, this boils down to a cumulative normal distribution of the form $N(c \mid M_f, S_f)$, where M_f and S_f depend on the specific model. Thus, the general form of the likelihood of identifying a particular item i, with confidence c, for a given model f, is given by (M_f and S_f depend on the form of the decision variable (f) for each model):

$$\mathcal{L}_f(r = i, c \mid \theta, \mathbf{q}) = \int_{-\infty}^{\infty} n\,(w_i|\mu_{q[i]}, \sigma_{q[i]}) \prod_{j \neq i} N\left(w_i \mid \mu_{q[j]}, \sigma_{q[j]}\right) N(c \mid M_f, S_f) \quad dw_i$$

Model-specific terms

The model-specific term in our likelihoods is the term $P(f(\mathbf{x}) > c \mid w_i, w_i = \max(\mathbf{w}))$, which we show below can be approximated as $1 - N\left(c \mid M_f, S_f\right)$ for all models.

Dependencies of the decision variable

The challenge in specifying these likelihood functions completely lies in the fact that, $f(\mathbf{x})$ is not independent of $w_i = \max(\mathbf{w})$ for all but the most trivial $f(\cdot)$. The reason is that the conditional distributions of w_js, given that they are smaller than w_i, will follow a truncated normal distribution (truncated at an upper bound w_i). We know of no analytical solution to $P(f(\mathbf{x}) > c \mid w_i, w_i = \max(\mathbf{w}))$, so we will adopt normal approximations via the mean and variance of the truncated normal distribution.

Truncated Normal Approximation

The mean ($m()$) and variance ($v()$) of a normal variable (x) truncated at b (such that $x < b$) are functions of the mean (μ_x) and variance (σ_x^2) of x, and b:

$$\beta = \frac{b - \mu_x}{\sigma_x} \qquad Z = \frac{\phi(\beta)}{\Phi(\beta)}$$

$$m(\mu_x, \sigma_x, b) = \mu_x - \sigma_x Z$$

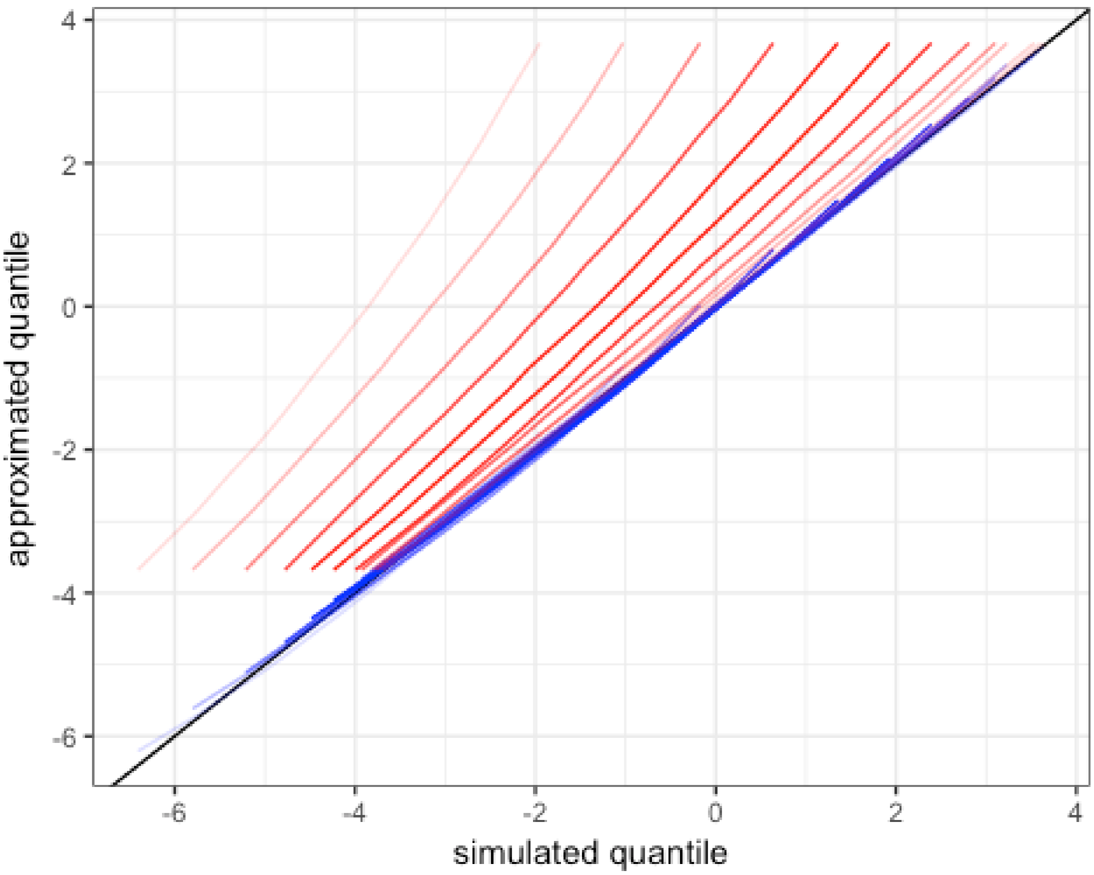
$$v(\mu_x, \sigma_x, b) = \sigma_x^2(1 - Z\beta - Z^2)$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the pdf and cdf of the standard normal distribution.

In the Best-Rest, Best-Ensemble, and Integration models, we approximate the distribution of $\sum_{j \neq i} w_j$ conditioned on $w_j < w_i$ as a normal distribution derived from the means and variances of the truncated w_j variables:

$$P\left(\sum_{j \neq i} w_j \mid w_j < w_i\right) \approx n\left(\sum_{j \neq i} m\left(\mu_{q[j]}, \sigma_{q[j]}, w_i\right), \sqrt{\sum_{j \neq i} v\left(\mu_{q[j]}, \sigma_{q[j]}, w_i\right)}\right)$$

This approximation is not perfect, given that the number of items on a given trial ($k \approx 6$) is quite small (so the central limit theorem should not be expected to hold), but is considerably better than simply ignoring the truncation. The graph below shows the quantile-quantile plots of this distribution, with the exact (numerically simulated) quantiles on the x-axis, and the approximation on the y-axis. Large deviations from the identity line (black), indicate a poor approximation. The blue lines (our truncated-normal approximation) do not deviate much from the identity line, while the red lines (corresponding to an approximation ignoring the truncation) are very far off (levels of transparency indicate the probability of a particular truncation $P(x) = \phi(x)\Phi(x)^{k-1}$.



The BEST (Independent Observations) model.

Under the BEST model (which we refer to as the Independent Observations model in the main text), the memory signal used for decision-making is simply the strongest memory signal:

$$f(\mathbf{x}) = \max(\mathbf{x})$$

Since we care about the value of this function when $w_i = \max\mathbf{w}$:

$$w_i = \max(\mathbf{w}) \Longrightarrow f(\mathbf{x}) = w_i + b$$

where b is the shared variance offset for that trial, with $b \sim n(0, \sigma_b)$. We are interested in evaluating $P(c < w_i + b)$. Given that b is normally distributed with mean 0, we can write this out in terms of the cumulative normal with mean w_i, and standard deviation σ_b:

$$P(f(\mathbf{x}) > c \mid w_i, w_i = \max(\mathbf{w})) = 1 - N(c \mid w_i, \sigma_b)$$

Or, in terms of our general expression $1 - N\left(c \mid M_f, S_f\right)$:

$$M_f = w_i$$
$$S_f = \sigma_b$$

The BEST-Rest model

Under the Best-Rest model, the decision variable is the difference between the memory signal of the strongest item on a given trial, and the average memory strength of the other items, but the correlated trial variance, b cancels out in the subtraction:

$$\begin{aligned} f(\mathbf{x}) &= \\ x_i - \frac{1}{k-1}\sum_{j\neq i} x_j &= \\ (w_i + b) - \frac{1}{k-1}\sum_{j\neq i}(w_j + b) &= \\ w_i - \frac{1}{k-1}\sum_{j\neq i} w_j & \end{aligned}$$

Conditioned on w_i being larger than all the w_js, the w_js will follow truncated normal distributions with means $m(\mu_{q[j]}, \sigma_{q[j]}, w_i)$ and variances $v(\mu_{q[j]}, \sigma_{q[j]}, w_i)$ (see the “truncated normal approximation” section). We will approximate the distribution of $\sum_{j\neq i} w_j$ (and thus $f(\mathbf{x})$) conditioned on $w_j < w_i$ as a normal distribution derived from the means and variances of the appropriate truncated normal distributions of w_j. Namely, the mean and variance of that sum, correpond to the sums of the means and variances of the components. Thus we get the expression:

$$P(f(\mathbf{x}) > c \mid w_i, w_i = \max(\mathbf{w})) \\ \approx 1 - N\left(c \mid w_i - \sum_{j\neq i}\frac{m(\mu_{q[j]}, \sigma_{q[j]}, w_i)}{k-1}, \sqrt{\sum_{j\neq i}\frac{v(\mu_{q[j]}, \sigma_{q[j]}, w_i)}{(k-1)^2}}\right)$$

Or, in terms of our general expression $1 - N\left(c \mid M_f, S_f\right)$:

$$M_f = w_i - \sum_{j\neq i}\frac{m(\mu_{q[j]}, \sigma_{q[j]}, w_i)}{k-1}$$

$$S_f = \sqrt{\sum_{j\neq i}\frac{v(\mu_{q[j]}, \sigma_{q[j]}, w_i)}{(k-1)^2}}$$

The Best-Ensemble model

Under the Best-Ensemble model (which we refer to as the Ensemble model in the main text):

$$
\begin{aligned}
f(\mathbf{x}) &= \\
x_i - \frac{1}{k}\sum_j x_j &= \\
(w_i + b) - \frac{1}{k}\sum_j (w_j + b) &= \\
w_i - w_i/k - \frac{1}{k}\sum_{j \neq i} w_j &
\end{aligned}
$$

Again, b is eliminated in the subtraction, and again we take the truncated normal approximation to the sum to approximate the conditional distribution of $f(\mathbf{x})$. Thus we approximate the conditional distribution of $f(\mathbf{x})$ as:

$$
\begin{aligned}
&P(f(\mathbf{x}) > c \mid w_i, w_i = \max(\mathbf{w})) \\
&\approx N\left(c \mid w_i(1 - 1/k) - \sum_{j \neq i} \frac{m(\mu_{q[j]}, \sigma_{q[j]}, w_i)}{k}, \sqrt{\sum_{j \neq i} \frac{v(\mu_{q[j]}, \sigma_{q[j]}, w_i)}{k^2}}\right)
\end{aligned}
$$

Or, in terms of our general expression $N(c \mid M_f, S_f)$:

$$
\begin{aligned}
M_f &= w_i(1 - 1/k) - \sum_{j \neq i} \frac{m(\mu_{q[j]}, \sigma_{q[j]}, w_i)}{k} \\
S_f &= \sqrt{\sum_{j \neq i} \frac{v(\mu_{q[j]}, \sigma_{q[j]}, w_i)}{k^2}}
\end{aligned}
$$

The Integration model

Under the Integration model, $f(\mathbf{X}) = \sum_j x_j$. We again factor out the influence of b and w on this sum (but in this case b does *not* cancel out, as there is no subtraction):

$$
\begin{aligned}
f(\mathbf{X}) &= \\
\sum_j x_j &= \\
\sum_j (w_j + b) &= \\
kb + \sum_j w_j &= \\
w_i + kb + \sum_{j \neq i} w_j &
\end{aligned}
$$

Since the sum over w and b are independent, their variances will add (the mean of b is 0, so it plays no role). We still need to take an approximation to the sum of the truncated w distributions. Together, this yields a conditional distribution of:

$$P(f(\mathbf{x}) > c \mid w_i, w_i = \max(\mathbf{w}))$$
$$\approx N\left(c \mid w_i + \sum_{j \neq i} m\,(\mu_{q[j]}, \sigma_{q[j]}, w_i),\ \sqrt{k^2\sigma_b^2 + \sum_{j \neq i} v\,(\mu_{q[j]}, \sigma_{q[j]}, w_i)}\right)$$

Or, in terms of our general expression $N(c \mid M_f, S_f)$:

$$M_f = w_i + \sum_{j \neq i} m\,(\mu_{q[j]}, \sigma_{q[j]}, w_i)$$
$$S_f = \sqrt{k^2\sigma_b^2 + \sum_{j \neq i} v\,(\mu_{q[j]}, \sigma_{q[j]}, w_i)}$$

Full likelihood model

The likelihood of identifying a particular item i, with confidence c, for a given model m, is given by:

$$\mathcal{L}_f(r = i, c \mid \theta, \mathbf{q}) =$$
$$\int_{-\infty}^{\infty} n\,(w_i | \mu_{q[i]}, \sigma_{q[i]})$$
$$\prod_{j \neq i} N\left(w_i \mid \mu_{q[j]}, \sigma_{q[j]}\right)$$
$$N(c \mid M_f, S_f) \quad dw_i$$

θ is a vector of parameters, consisting of
- σ_b: the standard deviation of the shared noise b
- μ_T, σ_T: the mean and standard deviation of w for targets
- σ_L: the standard deviation of w for the lures (the mean for lures is taken to be $\mu_L = 0$).

$\mathbf{q}$ is the trial specification: a vector of length k (the number of items), indicating whether or not each item $q[i]$ is a target (T) or a lure (L).
M_f and S_f depend on the form of the decision variable (f) for each model.

For all models, the probability that no item is identified is given simply as the probability of failing to identify any item at the lowest confidence level:

$$\mathcal{L}(r = \emptyset, c = \emptyset) = P(f(\mathbf{x}) < c_1) = 1 - \sum_i \mathcal{L}\,(r = i, c_1)$$

where c_1 corresponds to the lowest confidence level. This just reflects the assumption that for any item to be identified at any confidence level, the overall decision variable has to exceed the lowest confidence level.

Model	M_f	S_f
BEST	$M_f = w_i$	$S_f = \sigma_b$

BEST-REST	$M_f = w_i - \sum_{j\neq i} \frac{m(\mu_{q[j]}, \sigma_{q[j]}, w_i)}{k-1}$	$S_f = \sqrt{\sum_{j\neq i} \frac{v(\mu_{q[j]}, \sigma_{q[j]}, w_i)}{(k-1)^2}}$
BEST-ENSEMBLE	$M_f = w_i(1 - 1/k) - \sum_{j\neq i} \frac{m(\mu_{q[j]}, \sigma_{q[j]}, w_i)}{k}$	$S_f = \sqrt{\sum_{j\neq i} \frac{v(\mu_{q[j]}, \sigma_{q[j]}, w_i)}{k^2}}$
INTEGRATION	$M_f = w_i + \sum_{j\neq i} m\,(\mu_{q[j]}, \sigma_{q[j]}, w_i)$	$S_f = \sqrt{k^2\sigma_b^2 + \sum_{j\neq i} v\,(\mu_{q[j]}, \sigma_{q[j]}, w_i)}$

Simplifications for special cases

Although the general expression above for the likelihood can be used directly, it is somewhat unwieldy in that it obscures the relationship between parameters (e.g., σ_T) as these are only used by indexing via $q[j]$. We can omit q from the likelihood by writing out special cases for each type of trial (target present or absent) and identification (target, lure, none).

Model-independent part

First, we can write out simpler expressions for the first two terms of the integrand that do not depend on the model:

$$\mathcal{U} = n(w_i|\mu_{q[i]}, \sigma_{q[i]}) \prod_{j\neq i} N\left(w_i \mid \mu_{q[j]}, \sigma_{q[j]}\right)$$

Target-present trial, target ID: In this case (by definition), $q[i] =$T, and for all $j \neq i$, $q[j] =$L, consequently:

$$\mathcal{U} = n(w_i|\mu_T, \sigma_T)\; N(w_i \mid 0, \sigma_L)^{k-1}$$

Target-present trial, lure ID: In this case, $q[i] =$L, and $q[j] =$T for one j, and L for the rest. Critically, because there are $k-1$ lures, we have to account for all possible lures that might be identified. Thus:

$$\mathcal{U} = (k-1)\; n(w_i|0, \sigma_L)\; N(w_i \mid \mu_T, \sigma_T)\; N(w_i \mid 0, \sigma_L)^{k-2}$$

Target-absent trial, lure ID: In this case, $q[i] =$L for all i, but all k are equivalent, thus:

$$\mathcal{U} = k\; n(w_i|0, \sigma_L)\; N(w_i \mid 0, \sigma_L)^{k-1}$$

Model-dependent part

The $N\left(c \mid M_f, S_f\right)$ term of our likelihood (namely M_f and S_f), depend on the model, and the types of items present on the trial that were *not* identified. Consequently, we can simplify them to omit q by considering the two scenarios in which the non-identified items are all lures (target-present target ID, or target-absent lure ID), and those in which the non-identified items contain one target (target-absent lure ID).

For the sake of conciseness, below we use the abbreviations (note that the functions $m(\cdot)$ and $v(\cdot)$ are defined in the "truncated normal approximation" section):

$$\begin{aligned} m_{L<i} &= m(\mu_L, \sigma_L, w_i) \\ m_{T<i} &= m(\mu_T, \sigma_T, w_i) \\ v_{L<i} &= v(\mu_L, \sigma_L, w_i) \\ v_{T<i} &= v(\mu_T, \sigma_T, w_i) \end{aligned}$$

Model	Target-present trial, target ID or Target-absent trial, lure ID	Target-present trial, lure ID
BEST	$M_f = w_i$ $S_f = \sigma_b$	$M_f = w_i$ $S_f = \sigma_b$
BEST-REST	$M_f = w_i - m_{L<i}$ $S_f = \sqrt{\dfrac{v_{L<i}}{(k-1)}}$	$M_f = w_i - \dfrac{(k-2)\, m_{L<i} + m_{T<i}}{k-1}$ $S_f = \sqrt{\dfrac{(k-2)\, v_{L<i} + v_{T<i}}{(k-1)^2}}$
BEST-ENSEMBLE	$M_f = (\, w_i - m_{L<i}\,)(1 - 1/k$ $S_f = \sqrt{\dfrac{(k-1)v_{L<}}{k^2}}$	$M_f = \dfrac{w_i(k-1) - m_{L<i}(k-2) - m_{T<}}{k}$ $S_f = \sqrt{\dfrac{v_{T<i} + (k-2)v_{L<}}{k^2}}$
INTEGRATION	$M_f = w_i + (k-1)m_{L<i}$ $S_f = \sqrt{k^2\sigma_b^2 + (k-1)v_{L<i}}$	$M_f = w_i + m_{T<i} + (k-2)m_{L<i}$ $S_f = \sqrt{k^2\sigma_b^2 + v_{T<i} + (k-2)v_{L<i}}$

Constructing permutations.

By combining the model and trial specific special case terms, we can write out all the special case likelihoods by substituting the appropriate $\mathcal{U}$, M_f, and S_f terms into the expression below (note that here, $r \in \{\mathrm{T}, \mathrm{L}, \emptyset\}$, rather than the index of the identified item –i):

$$\mathcal{L}_f(r, c \mid \theta, \mathbf{q}) = \int_{-\infty}^{\infty} \quad \mathcal{U} \quad N\left(c \mid M_f, S_f\right) \quad dw_i$$

So, for instance, if we were defining the likelihood of identifying a lure, on a target-present trial ($\mathcal{L}_f(r = \mathrm{L}, c \mid \theta, \mathbf{q} = \{\mathrm{T}, \mathrm{L}, \ldots, \mathrm{L}\})$), under the BEST-ENSEMBLE model, we take:

$$\mathcal{U} = (k-1)\; n(w_i|0, \sigma_L)\; N(w_i \mid \mu_T, \sigma_T)\; N(w_i \mid 0, \sigma_L)^{k-2}$$

$$M_f = \frac{w_i(k-1) - m_{L<i}(k-2) - m_{T<i}}{k}$$

$$S_f = \sqrt{\frac{v_{T<i} + (k-2)v_{L<i}}{k^2}}$$

Yielding a complete likelihood of:

$$\begin{aligned}
\mathcal{L}_f(r = \mathrm{L}, c \mid \theta, \mathbf{q} = \{\mathrm{T}, \mathrm{L}, \ldots, \mathrm{L}\}) = \\
\int_{-\infty}^{\infty} (\,k-1)\; n(w|0, \sigma_L) \\
N(w_i \mid \mu_T, \sigma_T)\; N(w_i \mid 0, \sigma_L)^{k-2} \\
N(c \mid \frac{w_i(k-1) - m_{L<i}(k-2) - m_{T<i}}{k}, \\
\sqrt{\frac{v_{T<i} + (k-2)v_{L<i}}{k^2}})\quad dw_i
\end{aligned}$$

All simplifications can be assembled in this manner for a given model (target, lure id on target-present trials, and lure id on target-absent trials); however, we will not write them all out here.