Modeling face similarity in police lineups

Shen, Kyros; Colloff, Melissa; Vul, Edward; Wilson, Brent; Wixted, John T

DOI:
10.1037/rev0000408

Document Version
Peer reviewed version

Citation for published version (Harvard):
Shen, K, Colloff, M, Vul, E, Wilson, B & Wixted, JT 2022, 'Modeling face similarity in police lineups', *Psychological Review*. https://doi.org/10.1037/rev0000408

Modeling Face Similarity in Police Lineups

Kyros J. Shen[1], Melissa F. Colloff[2], Edward Vul[1], Brent M. Wilson[1], & John T. Wixted[1]

[1]Department of Psychology, University of California, San Diego

[2]Department of Psychology, University of Birmingham

Author Note

CSV files of the previously published data by Colloff, Wilson, Seale-Carlisle, & Wixted (2021) analyzed here have been deposited in OSF; see https://osf.io/uzk48/ (Colloff, 2021). CSV files of the new data reported here have also been deposited in OSF; see https://osf.io/fr4xd (Shen, 2022). Colloff et al. (2021) study design and hypotheses were preregistered; see https://osf.io/s4fq6. The new study was not preregistered. The data from Colloff et al. (2021) were presented at the 2019 meeting of the Psychonomic Society, and the models tested here were presented at the 2021 meeting of the Psychonomic Society.

Brent M. Wilson is now at the University of California, Los Angeles.

Correspondence concerning this article should be addressed to John T. Wixted (jwixted@ucsd.edu).

Abstract

Police investigators worldwide use lineups to test an eyewitness's memory of a perpetrator. A typical lineup consists of one suspect (who is innocent or guilty) plus five or more fillers who resemble the suspect and who all known to be innocent. Although eyewitness identification decisions were once biased by police pressure and poorly constructed lineups, decades of social-science research led to the development of reformed lineup procedures that provide a more objective test memory. Under these improved testing conditions, cognitive models of memory can be used to better understand and ideally enhance eyewitness identification performance. In this regard, one question that has bedeviled the field for decades is how similar the lineup fillers should be to the suspect to optimize performance. Here, we model the effects of manipulating filler similarity to better understand why such manipulations have the intriguing effects they do. Our findings suggest that witnesses rely on a decision variable consisting of the degree to which the memory signal for a particular face in the lineup stands out relative to the crowd of memory signals generated by the set of faces in the lineup. The use of that decision variable helps to explain why discriminability is maximized by choosing fillers that match the suspect on basic facial features typically described by the eyewitness (e.g., age, race, gender, etc.) but who otherwise are maximally *dissimilar* to the suspect.

Keywords: Eyewitness Identification, Filler Similarity, Feature Matching; Signal Detection Theory

Modeling Face Similarity in Police Lineups

Eyewitness identification (ID) tests are frequently used by police to identify the perpetrator of a crime. In years gone by, the outcomes of such tests were often inappropriately influenced by the officer administering the identification procedure. For example, the lineup administrator might lead the witness to believe that the suspect they are being asked to identify is already known to be guilty, thereby biasing the witness to make an ID even in the absence of a strong memory-match signal. Over the years, research designed to eliminate such interpersonal biases has led to science-based recommendations that, when followed, result in a lineup procedure that provides an objective test of the eyewitness's memory (National Research Council, 2014; Wells et al., 2020).

Perhaps because the social psychology of eyewitness identification is now well developed, cognitive theories have been increasingly used to conceptualize and measure lineup performance (Bull-Kovera & Evelo, 2021). This includes the formal cognitive modeling of eyewitness identification, which was pioneered some time ago by Clark (2003, 2008) and is now commonplace (e.g., Cohen et al., 2020; Colloff et al., 2021; Dunn et al., 2022; Kellen & McAdoo, in press; Lee & Penrod, 2019; Rotello & Chen, 2016; Starns et al., 2021; Wixted et al., 2018). Most of the recent modeling work in this domain is grounded in signal detection theory, and our focus here is on the three leading signal detection models of eyewitness identification. Because our focus is on theory, our analysis presupposes the proper testing conditions that have been worked out by social psychologists over the years. How often the police arrange such conditions in the real world is a separate issue that we do not address here.

The composition of police lineups

Live lineups were once the norm, but they have been largely replaced by photo lineups in many countries (e.g., the US, Germany, Australia; Fitzgerald et al., 2018). A proper photo lineup consists of one suspect and five or more physically similar fillers (Wells et al., 2020). The guilt or innocence of the suspect is unknown in actual police photo lineups (indeed, the lineup is being administered to help determine if the suspect is innocent or guilty), but the fillers are known to be innocent. Unlike the real-life scenario, researchers who investigate lineups in mock-crime studies control whether the suspect in a lineup is innocent or guilty. In a typical mock crime study, participants first watch a video of a perpetrator committing a simulated crime and are then presented with a photo lineup. A "target-present" (TP) lineup contains a photo of the guilty suspect surrounded by fillers, whereas a "target-absent" (TA) lineup contains a photo of an innocent suspect surrounded by fillers. The witness can either identify someone from the lineup (the suspect or a filler) or reject the lineup. Typically, that decision ends the experiment for a given participant, unlike a typical cognitive psychology experiment in which a participant's memory is tested multiple times. In studies of eyewitness identification, participants are typically tested only once because, in the real world, witnesses are typically tested only once.

The suspect is the only person of interest in a lineup, but the importance of the fillers should not be overlooked. Without the fillers, the eyewitness identification procedure reduces to a "showup" in which the singular suspect (innocent or guilty) is presented to the witness for a yes/no recognition decision. Studies have consistently found that lineups enhance discriminability compared to showups (e.g., Akan et al., 2020; Neuschatz et al., 2016; Wetmore et al., 2015; Wooten et al., 2020). Empirically, enhanced discriminability means that witnesses are better able to correctly sort suspects into the appropriate category (innocent or guilty).

Theoretically, enhanced discriminability means that the memory signals generated by innocent and guilty suspects overlap to a lesser degree (Wixted & Mickes, 2018). Empirical and theoretical discriminability usually go hand in hand (i.e., enhancing one usually enhances the other as well), and that has been found to be true of showups vs. lineups (e.g., Akan et al., 2020; Colloff & Wixted, 2020).

Although adding fillers to a showup, thereby creating a lineup, enhances discriminability, it is important to appreciate that not just *any* fillers would have that effect. Instead, the fillers need to be selected in such a way that the suspect does not gratuitously stand out. In other words, the lineup must be fair (e.g., Colloff et al., 2016; Lindsay & Wells, 1980). Two different approaches have been used to create fair lineups.

The first approach, which is the one most often used by the police (Police Executive Research Forum, 2013) and is sometimes used by researchers, is to select fillers whose faces appear to be similar to the suspect's face. But how similar should they be? Quite a few police departments choose fillers who are as similar to the suspect as possible to ensure a fair procedure (Police Executive Research Forum, 2013). However, as has long been appreciated, choosing fillers who are too similar would create an impossibly difficult lineup test (Wells, Rydell, & Seelau, 1993). Yet going in the other direction and making them too dissimilar (e.g., a heavily tattooed male suspect surrounded by non-tattooed female fillers) would result in an unfair lineup. Is there an *optimal* level of filler similarity? It is hard to answer that question without first considering the second approach to creating a fair lineup.

The second approach is to select fillers based on the description of the perpetrator provided by the eyewitness (Luus & Wells, 1991). This is the approach that is most often used by researchers and is sometimes used by the police (Police Executive Research Forum, 2013). As

an example, if the witness described the perpetrator as a clean-shaven white male in his 20s or 30s, and if the police locate a suspect who also fits that description, then fillers would be selected only if they, too, fit that description. The use of a description-matched approach avoids both the high and low extremes of filler similarity and results in levels of similarity characterized by "propitious heterogeneity" (Wells et al., 1993).

In addition to having propitious heterogeneity, is the average degree of filler similarity achieved using this description-matched approach also the optimal level of similarity? The answer to that question has remained a mystery for decades (e.g., Wells et al., 2020), and it might remain that way absent a formal cognitive model of filler similarity. Here, we describe and test competing signal detection models implemented in a feature-matching framework to (a) conceptualize the optimal level of filler similarity and (b) identify the underlying decision variable that witnesses use to decide whether or not to identify someone from the lineup. As described next, the optimal level of filler similarity is achieved by combining the two approaches described above, but with a twist.

Prior research on filler similarity

To protect the innocent, intuition suggests that fillers should be similar to the suspect. However, as noted above, Wells et al. (1993) cautioned against the use fillers that were too similar to the suspect because, in the extreme, it would make the task too difficult. They counterintuitively suggested that choosing *dissimilar* fillers would not be problematic so long as they were matched to the description of the perpetrator. To investigate the effect of manipulating filler similarity, they compared lineups comprised of description-matched fillers that were chosen from a large pool to be either maximally similar or maximally dissimilar to the suspect. The hit rate (proportion of guilty suspects identified from TP lineups) was lower when similar

fillers were used, but the false alarm rate (proportion of innocent suspects identified from TA lineups) did not differ significantly between the two conditions. This study did not include a pure description-matched condition, with fillers chosen from the pool without regard for similarity (i.e., an intermediate-similarity condition), but it seems reasonable to suppose that an intermediate hit rate would have been observed had such a condition been included. If so, it would mean that the best approach to choosing fillers would be to maximize dissimilarity between the description-matched fillers and the suspect. However, no theoretical mechanism was proposed to support that possibility, which may explain why the results of this seminal study have been almost universally understood to support the use of description-matched fillers only, not to support taking the additional step of maximizing dissimilarity to the suspect.

More recent studies have further documented the detrimental effect of choosing similar fillers relative to a condition involving an intermediate level of similarity. Fitzgerald et al. (2015) manipulated filler similarity to the suspect (ranging from medium to high) by morphing the suspect's photo in TA and TP lineups with the fillers. They found a significant decrease in the hit rate and a slight but non-significant decrease in the false alarm rate as filler similarity increased, supporting the idea that increasing filler similarity is potentially problematic. Oriet and Fitzgerald (2018) manipulated filler similarity to the suspect (again, medium to high) based on independent ratings of how similar the fillers were to the innocent or guilty suspect and reported comparable results. Carlson et al. (2019) compared lineups with description-matched fillers vs. lineups with fillers chosen because of their similarity to the suspect's face. Receiver operating characteristic (ROC) analysis showed that discriminability was higher when the fillers were intermediate in similarity (i.e., when they were simply matched to the description of the perpetrator without otherwise considering their similarity to the suspect).

Taken together, these studies indicate that, as Wells et al. (1993) argued long ago, one can go too far in selecting fillers who are similar to the suspect in an effort to create fair lineups. Doing so makes it harder for eyewitnesses to identify the guilty suspect while offering little or no protection to innocent suspects (beyond that already provided using description-matched fillers). In addition, the approach taken by Carlson et al. (2019) highlights an important consideration that has not often been addressed in prior inquiries into the optimal level of filler similarity: what measure of lineup performance, exactly, would be maximized if the optimal level of filler similarity were achieved? According to one reasonable definition, the optimal level of filler similarity is the one that maximizes the ability of eyewitnesses to discriminate innocent from guilty suspects, as measured by the area under the ROC.

In a recent study of filler similarity that we consider in detail here, ~10,000 once-tested participants watched a mock-crime video and were then presented with either a TA or TP lineup involving description-matched fillers who had low, medium, or high similarity to the suspect in the lineup. Prior to running the experiment, a large pool of fillers who matched the basic description of the perpetrator in the video was created. Using a separate group of participants, each filler was independently rated for similarity to the perpetrator using a 1-to-7 scale, and the median-similarity filler was designated to serve as the innocent suspect. The remaining fillers were then rated for similarity to that innocent suspect, again using a 1-to-7 scale. The similarity ratings to the guilty suspect and, separately, to the innocent suspect were used to manipulate filler similarity to the suspect over three levels (low, medium, or high) in the TP and TA lineups, respectively. Note that the low-similarity condition involved maximizing the *dissimilarity* of description-matched fillers to the suspect, as Wells et al. (1993) did.

For a given TP or TA lineup, the participant could identify someone from the lineup (the suspect or a filler) or reject the lineup. If someone was identified, a confidence rating was collected using a 100-point scale. In actual police lineups, filler IDs are inconsequential for the identified individual because fillers are known to be innocent. By contrast, suspect IDs are consequential for the identified individual because an identified suspect (innocent or guilty) may be tried, convicted, and sentenced to prison. To specify the optimal level of filler similarity, we therefore focus on the empirical effect of manipulating filler similarity on the consequential suspect ID hit and false alarm rates. However, because a model of eyewitness identification performance should be able to account for the full range of outcomes, the models we consider later are fit to suspect IDs, filler IDs, and lineup rejections from both TA and TP lineups.

Colloff et al. (2021) found that the use of low-similarity description-matched fillers maximized the hit rate to guilty suspects without measurably affecting the false alarm rate to innocent suspects (Figure 1), consistent with Wells et al. (1993). Moreover, ROC analysis indicated that empirical discriminability (the ability to sort innocent and guilty suspects into their correct categories) increased as filler similarity decreased (Figure 2).

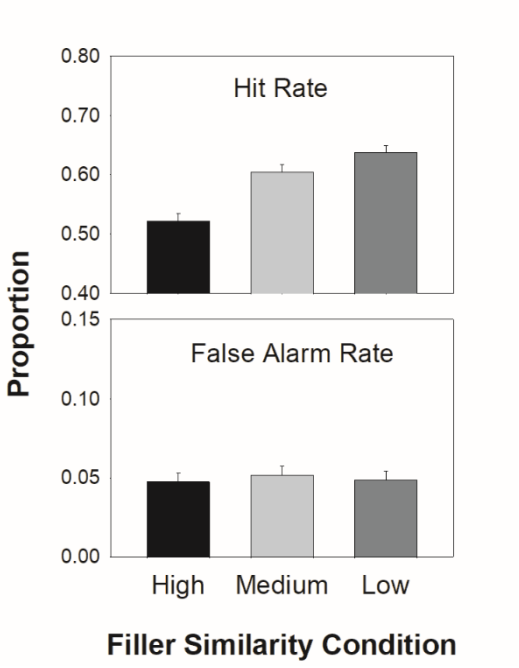


Figure 1. Hit and false alarm rates (i.e., correct and false suspect ID rates from TP and TA lineups, respectively) across three filler-similarity conditions in Experiment 1 of Colloff et al. (2021). All of the fillers matched the basic description of the perpetrator, but they varied in how similar they were to the (innocent or guilty) suspect in the lineup. The hit rate in the low-similarity condition was significantly higher than the hit rate in both the medium-similarity condition ($p = 0.043$) and high-similarity condition ($p < 0.001$). The hit rate in the medium-similarity condition was also higher than that of the high-similarity condition ($p < 0.001$). The corresponding comparisons for the false-alarm rates did not approach significance (p = 0.726, 0.874, and 0.610, respectively).

For those familiar with ROC analysis, eyewitness identification ROCs seem unusual at first glance. For a simple yes/no recognition task, generating ROC data would yield a familiar curve, with hit and false alarm rates both ranging from 0 to 1. However, for a lineup, the hit rate generally ranges from 0 to a maximum of less than 1.0. The reason is that even when the most liberal decision criterion is used, in which case a TP lineup would never be rejected, the guilty suspect will not always yield the strongest signal, in which case a filler will be identified. The same is true for the innocent suspect. In a fair lineup in which the innocent suspect does not stand out, the maximum false alarm rate is 1 / *k*, where *k* is lineup size. As an example, the maximum false alarm rate for a typical 6-person photo lineup is 1 / 6 = .167, and a typical overall false alarm rate is ~.05. Despite its somewhat unfamiliar appearance, the *partial* ROC (*p*AUC)

provides a measure of discriminability at the level of observable data (Gronlund et al., 2014; Robin et al., 2011).

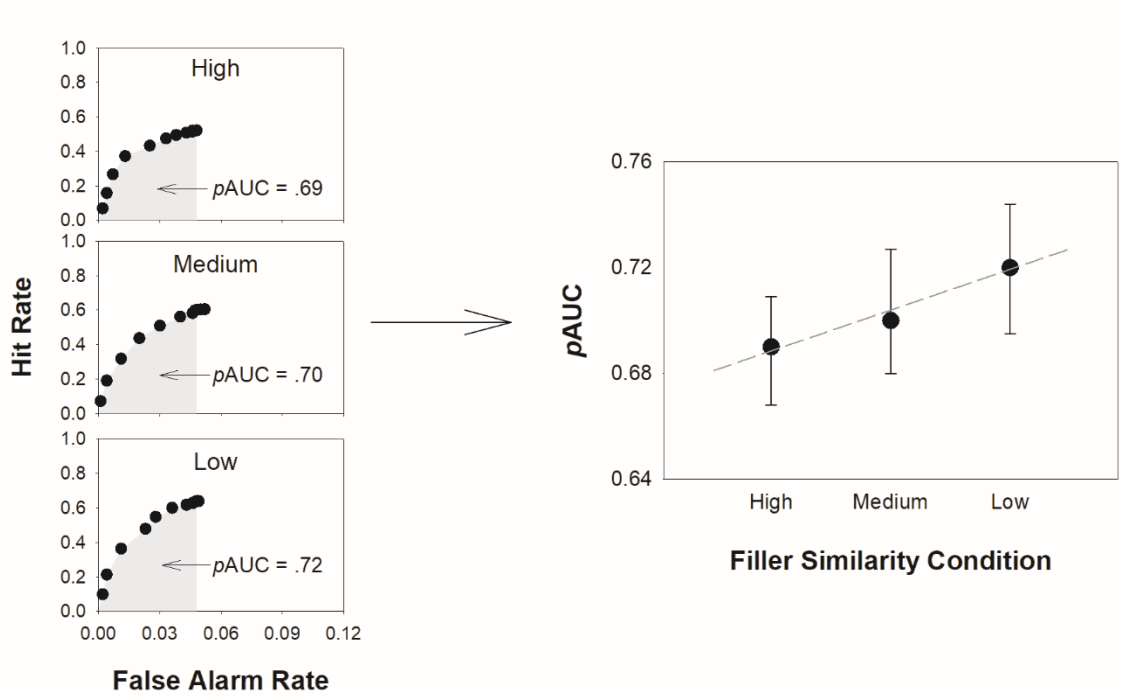


Figure 2. Left panel: Receiver operating characteristic curves for the three filler-similarity conditions (High, Medium, or Low) of the Suspect Similarity experiment reported by Colloff et al. (2021). Discriminability is quantified by the partial area under the curve (*p*AUC) represented by the shaded regions, using a common false-alarm rate across the three conditions (.048). The *p*AUC values have been corrected to fall on a scale ranging from .50 (chance) to 1.0 (perfect discriminability). The low-similarity *p*AUC was significantly larger than the high-similarity *p*AUC ($p = 0.029$, one-tailed, per preregistration), but the medium-similarity *p*AUC did not differ significantly from either of the other two conditions. Right panel: the same *p*AUC data potted with bootstrapped 95% confidence intervals. Despite having only three points, the slope of a line fit to the *p*AUC values (i.e., the slope of the dashed line) was significantly greater than 0, $t(1) = 19.82$, $p = .032$ (two-tailed).

The ROC results reported by Colloff et al. (2021) provide an answer to a longstanding question: what is the optimal level of filler similarity in a police lineup? The optimal level of filler similarity is not achieved by matching fillers to the witness's description of the perpetrator (the standard recommendation). Instead, it is achieved using a two-step process: step one is to create a pool of description-matched fillers (all of whom would be suitable for creating a fair

lineup), and step two is to *maximize dissimilarity* by choosing from that pool the fillers who are the least similar to the suspect in the lineup (Wells et al., 1993).

What do these results tell us about how eyewitnesses make decisions from the memory signals generated by the faces in a lineup? Addressing that question is our main focus, and competing signal detection models provide different answers. For example, one possibility is that witnesses first locate the face in the lineup that generates the strongest memory-match signal (the MAX face) and then base their decision and their confidence on the magnitude of that memory signal alone (Clark et al., 2011; Sauer et al., 2008). In other words, the signal associated with the MAX face, considered in isolation, would be the decision variable. This signal detection model is known as the Independent Observations model. Another possibility is that witnesses instead compute the *difference* between the signal associated with the MAX face vs. the average memory-match signal associated with other faces in the lineup and then base their decision and confidence on the magnitude of that difference score. This signal detection model is known as the Ensemble model.[1] Still a third possibility is that witnesses compute the *sum* of the memory signals associated with the faces in the lineup and then decide whether or not to choose the MAX face based on the magnitude of that summed score. This signal detection model is known as the Integration model.

These are not the only possible signal detection models of eyewitness identifications from a lineup, but they are the leading contenders. Prior research weighs against the Integration model but does not conclusively favor either the Independent Observations model or the Ensemble model (Wixted et al., 2018). However, manipulating filler similarity turns out to provide a more decisive test of their predictions. As noted by Colloff et al. (2021), different degrees of filler

[1] The Ensemble model is linearly related to (and is therefore the mathematical equivalent of) a MAX minus rest decision rule (Clark, 2011, adapted from Sauer et al., 2008).

similarity can be conceptualized in terms of the number of facial features that match between the fillers and the suspect in a lineup. We pursue that idea here to (a) clarify what the competing signal detection models predict about the effect of manipulating filler similarity and (b) provide a concrete feature-based interpretation of their free parameters.

To illustrate the basic idea, consider a task that is simpler than a 6-person lineup but is similar to it in that also involves making a recognition decision about more than one test stimulus: two-alternative forced-choice (2AFC). In 2AFC, a test trial involves the presentation of two test items, an old (target) item and a new (foil) item, and the participant's task is to pick the target. In the equal-variance Gaussian signal detection framework, discriminability (d') for a 2AFC task is equal to $d'_{2AFC} = \frac{\sqrt{2}(\mu_{Target} - \mu_{Foil})}{\sigma\sqrt{(1-\rho)}}$, where μ_{Target} represents the mean memory signal associated with previously presented old items, μ_{Foil} represents the mean memory signal associated with new items, σ represents the common standard deviation of the target and foil distributions, and ρ represents the degree to which the target and foil memory signals are correlated, as they would be if they were similar to each other (Wixted, 2020). For example, after a participant has studied a list of common objects, if recognition test trial 1 consisted of a target violin paired with a foil violin, and recognition test trial 2 consisted of a target apple paired with a foil apple, and so on, the pair of memory signals on each trial would be correlated because the two test items would share many features.

To say that the memory signals would be correlated means that if the memory signal for the target happened to be weak, then the memory signal for the corresponding foil would also be weak, but if the memory signal for the target happened to be strong, then the memory signal for the corresponding foil would also be strong. All else equal, correlated memory signals enhance discriminability on a 2AFC task relative to uncorrelated signals (e.g., Hintzman, 1988) in much

the same way that correlated dependent measures in a within-subject design enhance statistical power compared to the uncorrelated measures obtained using a between-subject design.

The parameters of the 2AFC equation—like the parameters of the similar equations for the lineup models that we consider later—will change depending on the similarity between the foils and targets. For example, as the foils all become increasingly similar to their corresponding targets (i.e., as they share more and more features), μ_{Foil} will approach μ_{Target}. Therefore, the numerator of the d'_{2AFC} equation presented above, $\sqrt{2}(\mu_{Target} - \mu_{Foil})$, will decrease, exerting downward pressure on discriminability. At the same time, the correlation between targets and foils (ρ) will increase, so the denominator, $\sigma\sqrt{(1-\rho)}$, will also decrease, exerting upward pressure on discriminability. But how fast will the opposing forces in the numerator and denominator change with respect to each other as a function of filler similarity? By specifying each of these parameters in terms of the number of features that match between the simultaneously presented test items and also between those items and memory of the target, it becomes possible to precisely predict what the effect of manipulating similarity should be (separately for each model of eyewitness identification).

A general feature-matching model of eyewitness identification

To be clear, our goal is not to propose a new simulation-based feature-matching model of eyewitness identification, such as the WITNESS model (Clark, 2003). Instead, our goal is to use feature-matching logic to derive the extant signal detection models of eyewitness identification (the Independent Observations model, the Ensemble model, and the Integration model) from the ground up. The models we later fit to the empirical filler-similarity data reported by Colloff et al. (2021) and to new filler-similarity data reported here will be the same models that have been previously fit to other data (Wixted et al., 2018). However, having a feature-matching version of

each model will allow us to formally specify their predictions about the effect of manipulating filler similarity while providing a feature-based interpretation of their key parameters.

Overview and Context

As noted earlier, a typical lineup recognition memory experiment in the field of eyewitness identification differs in several important ways from a typical old/new recognition memory test in the field of cognitive psychology. For example, an old/new recognition memory experiment in cognitive psychology typically involves a study list, followed by the presentation of test items, one at a time, each for an old/new "detection" decision. By contrast, in a lineup recognition memory experiment in the field of eyewitness identification, the "list" typically involves the single face of the perpetrator seen in a mock-crime video, with multiple test stimuli presented simultaneously in a lineup for a detection ("is the perpetrator's face in the lineup?") plus identification ("if so, which face is it?") decision.

Models of list memory in cognitive psychology often assume a global matching process according to which a given test stimulus is compared to the memory representation of each list item (Gillund & Shiffrin, 1984; Nosofsky, 1991; Osth & Dennis, 2015; Shiffrin & Steyvers, 1999). For example, Nosofsky (1991) presented participants with a study list of 10 similar schematic faces. On a subsequent old/new recognition test, a given test face, i, was theoretically compared to each of the $j = 1 \rightarrow 10$ studied faces in memory in a 2-step process. First, the psychological distance between test face i and memorized face j (d_{ij}) was computed (e.g., the Euclidean distance between their multidimensional representations), and second, a corresponding similarity score (s_{ij}) was computed, where $s_{ij} = e^{-d_{ij}}$ (Shepard, 1958, 1987). Theoretically, the degree to which test face i activates the memory representation of face j (a_{ij}) is given by $a_{ij} = s_{ij} + e_j$, where the e_j values are independent and identically distributed normal random variables

with a mean 0 and variance of 1. Once these values are determined, the overall activation (A_i)—that is, the familiarity of test face i—is determined by the sum of the similarity-based activations over the 10 faces in memory: $A_i = \sum_{j=1}^{10} a_{ij}$. This is a *summed similarity* model of recognition memory.

Here, we take an analogous summed similarity approach to conceptualize eyewitness identification except that, in our approach, similarity is evaluated at the level of individual facial features (cf. Lacroix, et al., 2006). In essence, we conceptualize the discrete features of the singular perpetrator's face in memory as if they were a list of items (e.g., blue eyes, thick eyebrows, square chin, etc.). On a subsequent face recognition test, the mean strength of the memory-match signal for a given feature of the test face is determined by its degree of similarity to the corresponding feature of the perpetrator's face in memory. A similarity score is computed separately for each feature, and the scores are then summed to yield an overall similarity score for the test face. We assume that only corresponding features are relevant for computing feature-level similarity scores because a global matching process in which each feature is compared to all of the facial features of the perpetrator's face in memory would theoretically return a similarity score of ~0 for all of the non-corresponding features (because, for example, eye color and mouth shape are not at all similar).

As an example, if d_{ij} is a hypothetical distance measure in psychological space between (a) feature i of lineup face j and (b) the corresponding feature of the perpetrator's face in memory, then the feature-level similarity (s_{ij}) would be $s_{ij} = e^{-\alpha d_{ij}}$, where α is a scaling constant. If the two features being compared happen to match (in which case $d_{ij} = 0$), then $s_{ij} = 1$, but if they do not match, the similarity score would fall between 0 and 1, and its magnitude would depend on the psychological distance separating the two corresponding features (e.g., blue

eyes vs. green eyes). As the exponential scaling constant α becomes large, the s_{ij} similarity scores approach a binary match-vs.-mismatch (1 vs. 0) distribution, and we adopt that simplifying assumption throughout (i.e., we assume large α). Lacroix, et al. (2006) used an analogous binary scheme in their feature-based summed similarity model. Relaxing this assumption by allowing for a more gradual exponential decay as a function of psychological distance would not change the essence of our theory in any significant way.

The activation signal for a given feature (a_{ij}) is given by $a_{ij} = s_{ij} + e_i$, where, s_{ij} equals 1 or 0 (old target vs. new foil) depending on whether or not it matches the corresponding feature in memory, and the e_i scores for different features are independent identically distributed normal random variables with mean of 0 and variance of 1 (Figure 3A). Like Nosofsky (1991), we assume that there are many sources of noise, and we take the existence of feature-level Gaussian error for granted rather than explicitly modeling it.

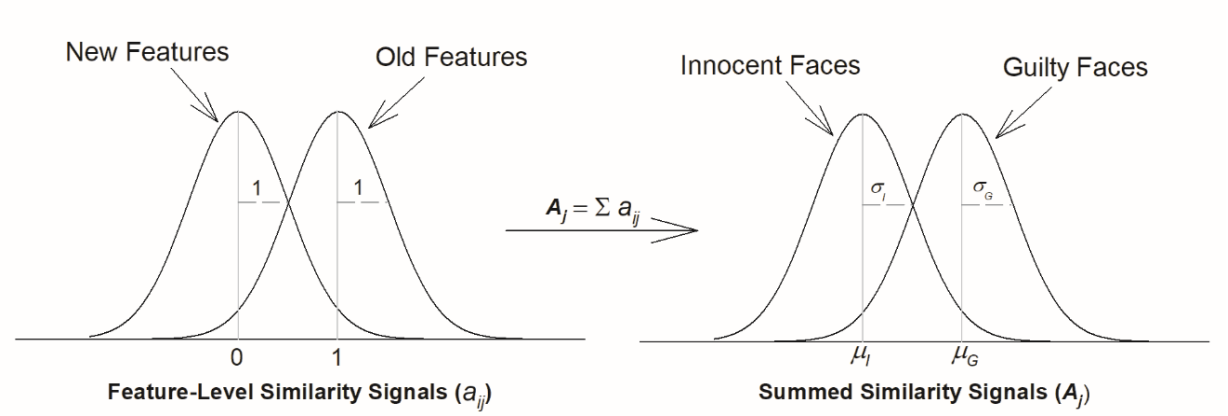


Figure 3. A. An equal-variance signal detection model of the similarity-based activations (a_{ij}) generated by the individual features (indexed by *i*) of face *j*, depending on whether they match memory of the perpetrator (old features) or not (new features). For convenience, we set $u_{New} = 0$, $u_{Old} = 1$, and $\sigma_{New} = \sigma_{Old} = 1$. B. An equal-variance signal detection model of the summed similarity-based activations (A_j) for faces (indexed by *j*), depending on whether the faces are innocent (mean and standard deviation = μ_I and σ_I, respectively) or guilty (mean and standard deviation = μ_G and σ_G, respectively). Innocent faces are composed of mostly new features, whereas guilty faces are composed of mostly (or only, in the simplest case) old features.

The overall activation (i.e., familiarity) signal for face j in a lineup, A_j, is given by $A_j = \sum_{i=1}^{n} a_{ij}$, where n is the number of features that define a face (note that this sum is over features of the perpetrator's face in memory, not over faces in the lineup). This summed similarity signal for the faces of guilty suspects across lineups is the "target" distribution at the level of faces, and the summed familiarity signal for the faces of innocent suspects (and innocent fillers) across lineups is the "foil" distribution at the level of faces (Figure 3B).

Critically, in a lineup, some features will be shared by everyone because, as a condition for being included in a description-matched lineup, every face must match the physical description of the perpetrator provided by the eyewitness. If a described feature like blue eyes is shared by all the faces in the lineup, and if that feature happens to generate a relatively strong memory match signal for one face in the lineup, then it will do so for the other faces as well. The fact that feature-level memory signals are shared across the faces in a lineup is why lineup memory signals are theoretically correlated.

In terms of this modeling framework, increasing filler similarity to the suspect (innocent or guilty) involves increasing the number of features that match across the faces in the lineup above the number of features that already match because the fillers were selected to fit the witness's description of the perpetrator. Increasing the number of features that match will increase the degree to which the face memory signals are correlated, whether or not those extra matching features also match memory of the perpetrator. Similarly, *decreasing* filler similarity to the suspect involves decreasing the number of features that happen to match across faces in a lineup, thereby decreasing the correlation, without changing the features that match by design (i.e., because it is a description-matched lineup).

Formal specification

We assume that an unfamiliar face is represented by *n* features (features $f_1 \rightarrow f_n$) and that each feature has v possible settings. This is similar to how faces are represented in the WITNESS model (Clark, 2003). In the concrete example we use throughout, $n = 20$ and $v = 5$. If feature 1 is race/ethnicity, for example, its $v = 5$ possible settings might be (1) = Caucasian, (2) = African American, (3) = Hispanic, (4) = Asian, and (5) = Pacific Islander. If feature 2 is eye shape, its possible settings might be (1) = Round eyes, (2) = Narrow eyes, (3) = Slanted eyes, (4) = Upturned eyes, and (5) = Downturned eyes. Thus, a Caucasian with upturned eyes would have settings of $f_1 = 1$ and $f_2 = 4$ (and so on through f_{20}).

Because a face is more than simply the sum of its parts (where a "part" is a low-level physical feature like eye color), higher-level information, such as relational/configural information (e.g., distance between the eyes and mouth) and inferred personality traits (e.g., trustworthiness) are also conceptualized as features for modeling purposes (cf. Cox & Shiffrin, 2017; Nelson & Shiffrin, 2013). The encoding of higher-level features would account for why a "composite face" consisting of the upper half of one previously seen face and the lower half of another previously seen face is less likely to be recognized as "old" than a fully intact previously seen face (e.g., Meltzer & Bartlett, 2019).

After witnessing a crime, we assume that the witness has seen and successfully encoded all *n* features of the perpetrator's face (whereas the WITNESS model allows for an error-prone encoding process), each with one of v possible settings. We further assume that a certain number of those feature settings (n_D) will be included in the description of the perpetrator provided by the eyewitness to the police. Although the number of features described surely varies from witness to witness, for simplicity, we fix n_D at 5.

For notational purposes, we use capital letters to represent the features of (a) the perpetrator's face in memory ($P_1 \rightarrow P_{20}$), (b) the guilty suspect's face in a TP lineup ($G_1 \rightarrow G_{20}$), (c) the innocent suspect's face in a TA lineup ($I_1 \rightarrow I_{20}$), and (d) a filler's face in either type of lineup ($F_1 \rightarrow F_{20}$). Of the $P_1 \rightarrow P_{20}$ feature settings of the perpetrator's face in memory, let settings $P_1 \rightarrow P_5$ correspond to the $n_D = 5$ features described to the police. As noted earlier, using a description-matched approach, photos are selected for inclusion in a lineup (suspect and fillers alike) precisely because they match these 5 features in the witness's description. Thus, the feature settings of everyone in a TA lineup or a TP lineup will not only match each other but will also match the settings in memory for $P_1 \rightarrow P_5$. That is, $G_1 \rightarrow G_5$ (if it is a TP lineup), $I_1 \rightarrow I_5$ (if it is a TA lineup), and $F_1 \rightarrow F_5$ (whether it is a TP or a TA lineup) will match the corresponding features settings of the perpetrator's face in memory ($P_1 \rightarrow P_5$). For these 5 features of every face in a TA or TP lineup, the memory-match signals they generate are drawn from the old feature distribution in Figure 3A (with a mean and standard deviation of 1).

Features $f_1 \rightarrow f_5$ are non-diagnostic of guilt because they are shared by everyone in a description-matched lineup (Wixted & Mickes, 2014). By contrast, features $f_6 \rightarrow f_{20}$ are potentially diagnostic of guilt because their settings for the guilty suspect's face are more likely to match memory than the corresponding settings for non-guilty lineup members. The non-guilty lineup members consist of innocent suspects in TA lineups and fillers in either TA or TP lineups.

Because we are modeling the ideal case of error-free encoding, we assume that the settings for the 15 potentially diagnostic features of the guilty suspect's face ($G_6 \rightarrow G_{20}$) are the same as the settings for the corresponding features of the perpetrator's face in memory ($P_6 \rightarrow P_{20}$). For example, if the eye color setting for the perpetrator is blue, the eye color setting for the guilty suspect is also blue. This means that not only $G_1 \rightarrow G_5$ (the description-matched

features) but also $G_6 \rightarrow G_{20}$ are the same as the settings for the corresponding features of the perpetrator's face in memory ($P_1 \rightarrow P_{20}$). Thus, the number of features on the guilty suspect's face that have the same feature setting as memory of the perpetrator (n_G) is $n_G = n = 20$. Although all 20 feature settings between P_i and G_i match, the strength of the noisy memory-match signals they generate differs across features because they are independently drawn from the feature-level old distribution (Figure 3A).

For non-guilty lineup members (innocent suspects, TA fillers, and TP fillers), the settings for $I_1 \rightarrow I_5$ and $F_1 \rightarrow F_5$ match the settings of $P_1 \rightarrow P_5$ by design, whereas the settings of the non-description-matched features, $I_6 \rightarrow I_{20}$ and $F_6 \rightarrow F_{20}$, will match the settings of $P_6 \rightarrow P_{20}$ in memory due to chance alone. Because each feature has v = 5 possible settings, the probability of a chance match (p) is $p = 1/v = 1/5 = .2$. Recall that the settings for $I_1 \rightarrow I_5$ and $F_1 \rightarrow F_5$ match the settings of $P_1 \rightarrow P_5$ in memory by design. Thus, on average, and assuming independence, the total number of features with settings that match the corresponding feature settings of the perpetrator's face in memory for the innocent suspect (n_I), a TA filler ($n_{F_{TA}}$), and a TP filler ($n_{F_{TP}}$) is given by $n_I = n_{F_{TA}} = n_{F_{TP}} = n_D + p(n - n_D) = 5 + .20(20 - 5) = 5 + 3 = 8$. This means that, for a given face, 8 feature-level memory-match signals are independently drawn from the old feature distribution and 12 are independently drawn from the new feature distribution (Figure 3A). The values of n_I, $n_{F_{TA}}$, and $n_{F_{TP}}$ are equal because they are all people who match the description of the perpetrator but who are not guilty (so they were not previously seen), but these values will not necessarily remain equivalent when filler similarity to the suspect is manipulated.

If, on average, 8 features of a non-guilty face match the corresponding features of the perpetrator's face in memory, then there would be variability in the number of features that that

match memory across different lineups. However, compared to the Gaussian variability that we assume, the additional variance contributed by this multinomial variability would be small. Moreover, including provisions for multinomial variability would add complexity without adding commensurate theoretical insight (so far as we can determine, at least), so we ignore it in our main analysis. In the appendix, we provide a simulation-based version of our account that allows for multinomial variability for every feature-based setting, and it yields conclusions that are consistent with the simpler (and mathematically tractable) version we pursue here.

The overall memory-match signal for a given face in the lineup (i.e., its similarity to the perpetrator's face in memory) is the sum of the *n* feature-level memory-match signals. Because we assume that $n_G = n$ (i.e., the feature settings on the face of the guilty suspect match all *n* feature settings of the perpetrator in memory), then across many lineups, the mean of the summed memory signal for guilty suspects (μ_G) would simply be $\mu_G = n_G = n$ (because each matching feature contributes a signal strength of 1, on average). In addition, variances sum, and the variance of each feature-level memory-match signal is 1. Thus, $\sigma^2 = n_G = n$ as well. In our running example, $n = 20$, so the mean of the summed memory signal for the guilty suspect is $\mu_G = 20$, and the standard deviation of the summed memory signal is $\sigma = \sqrt{20}$.[2]

As noted above, the average number of feature settings that match the settings of the perpetrator's face in memory for innocent suspects (n_I), TA fillers ($n_{F_{TA}}$), and TP fillers ($n_{F_{TP}}$) are all equal to $n_D + p(n - n_D)$. In our running example, this comes to 5 + .20(20 – 5) = 5 + 3 = 8 (i.e., 5 features match by design and 3 of the 15 potentially diagnostic features match by chance). It follows that the means of the innocent suspect distribution (μ_I), target-absent filler

[2] In terms of expected value (E), $\mu_G = E\left[\sum_1^{n_G} \mu_{target} + e_i\right] = \sum_1^{n_G} \mu_{target} + E[e_i]$, where $e_i \sim N(0,1)$. Because $E[e_i] = 0$ and $\mu_{target} = 1$, $\mu_G = \sum_1^{n_G} 1 + 0 = n_G = n$. In addition, $\sigma^2 = E\left[\sum_1^{n_G} e_i^2\right] = \sum_1^{n_G} E[e_i^2]$. Because $E[e_i^2] = 1$, $\sigma^2 = \sum_1^{n_G} 1 = n_G = n$. Thus, $\sigma = \sqrt{n}$.

distribution ($\mu_{F_{TA}}$), and target-present filler distribution ($\mu_{F_{TP}}$) are all equal to 8. However, because variances sum, the standard deviation in each case would still be $\sigma = \sqrt{n} = \sqrt{20}$.[3]

Consider the distribution of the overall (summed) memory signals for the faces of innocent and guilty suspects, as illustrated in Figure 4 (conceptually, this is the same model illustrated earlier using a more condensed format in Figure 3B). These face-level memory signals constitute the core signal detection model from which more specific models will be derived depending on what decision rule eyewitnesses are assumed to use. Figure 4 corresponds to the simplest face-memory test, which is a showup. In a showup, the witness is presented with a single face, either an innocent suspect or the guilty suspect (no fillers). In terms of underlying memory signals associated with innocent and guilty suspects, discriminability is represented by d'_{IG}, where $d'_{IG} = \frac{\mu_G - \mu_I}{\sigma}$. This discriminability measure will later play an important role in distinguishing between competing signal detection models of lineup performance because it is this measure that should be affected by manipulations of filler similarity in different ways (depending on which model is correct). Expressed in terms of the feature matching model, $d'_{IG} = \frac{n_G - n_I}{\sqrt{n}} = \frac{n - n_I}{\sqrt{n}}$. In our running example, where $n = 20$ and $n_I = 8$, $d'_{IG} = \frac{20-8}{\sqrt{20}} = 2.68$.

[3] That is, with $\hat{G}$ representing non-guilty lineup members, $\mu_{\hat{G}} = E\left[\sum_1^{n_{\hat{G}}} \mu_{target} + e_i\right] + E\left[\sum_1^{n-n_{\hat{G}}} \mu_{foil} + e_i\right] = E\left[\sum_1^{n_{\hat{G}}} 1\right] + E\left[\sum_1^{n_{\hat{G}}} 0\right] + E\left[\sum_1^{n-n_{\hat{G}}} 0\right] + E\left[\sum_1^{n-n_{\hat{G}}} 0\right] = E\left[\sum_1^{n_{\hat{G}}} 1\right] = n_{\hat{G}}$. Similarly, $\sigma^2 = E\left[\sum_1^{n_{\hat{G}}} e_i^2\right] + E\left[\sum_1^{n-n_{\hat{G}}} e_i^2\right] = \sum_1^{n_{\hat{G}}} E[e_i^2] + \sum_1^{n-n_{\hat{G}}} E[e_i^2] = \sum_1^{n_{\hat{G}}} 1 + \sum_1^{n-n_{\hat{G}}} 1 = n_{\hat{G}} + (n - n_{\hat{G}}) = n$, so $\sigma = \sqrt{n}$.

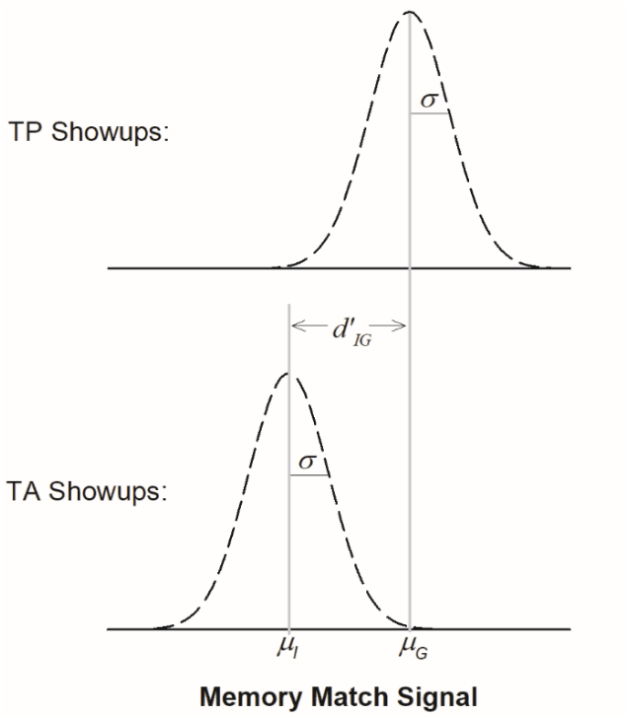


Figure 4. Face-level memory-match signals in a showup resulting from summing feature-level memory-match signals. d'_{IG} is the difference between the mean of the guilty suspect distribution (the face-level target distribution) and the mean of the innocent suspect distribution (the face-level foil distribution) in standard deviation units. That is, $d'_{IG} = \frac{\mu_G - \mu_I}{\sigma}$.

A lineup is simply a showup with description-matched fillers added to the procedure. When fillers are added, d'_{IG} remains a key measure, but additional discriminability measures now come into play. Figure 5 illustrates the lineup scenario when description-matched fillers are added to a showup. Note that in a TA lineup, the innocent suspect is effectively just another filler (i.e., someone who matches the description but did not commit the crime). This means that from the perspective of the witness, there is no distinction between an innocent suspect vs. a filler. Thus, the memory-strength distributions for the innocent suspect and the TA fillers are one and the same (bottom panel of Figure 5).

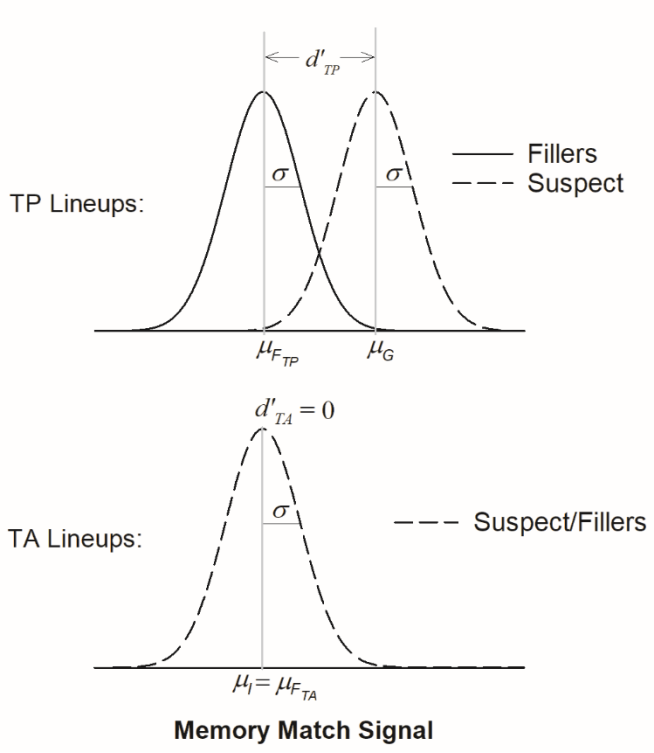


Figure 5. Face-level memory-match signals in a lineup resulting from summing feature-level memory-match signals. d'_{TP} is the difference between the mean of the TP filler distribution and the guilty suspect distribution in standard deviation units. That is, $d'_{TP} = \frac{\mu_G - \mu_{F_{TP}}}{\sigma}$. Similarly, d'_{TA} is the standardized difference between the TA filler distribution and the innocent suspect distribution. Because $\mu_I = \mu_{F_{TA}}$, $d'_{TA} = 0$.

As illustrated in the top panel of Figure 5, the separation between the underlying memory signals associated with guilty suspects and fillers within TP lineups is given by $d'_{TP} = \frac{\mu_G - \mu_{F_{TP}}}{\sigma}$. Similarly, as illustrated in the bottom panel of Figure 5, the separation between the underlying memory signals associated with innocent suspects and fillers within TA lineups is given by $d'_{TA} = \frac{\mu_I - \mu_{F_{TA}}}{\sigma}$ (Wixted et al. , 2018, 2021). In terms of the feature-matching model, $d'_{TP} = \frac{n_G - n_{F_{TP}}}{\sqrt{n}} = \frac{n - n_{F_{TP}}}{\sqrt{n}} = \frac{20-8}{\sqrt{20}} = 2.68$, and $d'_{TA} = \frac{n_I - n_{F_{TA}}}{\sqrt{n}} = \frac{8-8}{\sqrt{20}} = 0$. In a fair lineup, $d'_{TA} = 0$ because n_I (number of innocent suspect features that match memory of the perpetrator) is equal to $n_{F_{TA}}$ (number of TA filler features that match memory of the perpetrator). So far, $d'_{IG} = d'_{TP}$, but we will see that this is not always the case, so they must be considered separately.

Note that Figure 5 illustrates how memory signals are distributed in a lineup, but it is not yet a model of eyewitness identification performance because no decision criterion has been added. Moreover, a decision criterion cannot be added until the *decision variable* is specified, and specific signal detection models of lineup performance differ as to what they consider the decision variable to be. In addition, our analysis so far applies to uncorrelated memory signals, but as noted earlier, the memory signals generated by the faces in a lineup are correlated by design. They must be at least somewhat correlated because certain facial features of every lineup member are selected to match the witness's description of the perpetrator. Therefore, the competing signal detection models we consider later must include provisions for correlated memory signals. This means that the equations for each lineup model must include ρ as a parameter, just as the 2AFC equation presented earlier did. Before specifying the competing models in feature-matching terms, we first specify how the correlation parameter they have in common (i.e., ρ) can be conceptualized in feature-matching terms.

Correlated memory signals

As has long been known, positively correlated memory signals can enhance discriminability on a recognition memory test (e.g., Hall, 1979). As noted earlier, correlated memory signals can enhance discriminability for essentially the same reason that within-subject experimental designs can increase power compared to between-subject experimental designs. In a within-subject design, the dependent measures are usually correlated across conditions (e.g., if Subject 1 generates a relatively high score in Condition 1, that same subject is likely to generate a relatively high score in Condition 2). If the scores are in fact positively correlated, then computing a difference score for each subject would subtract out random error (e.g., effects on

the dependent measure arising from extraneous variables like age differences, gender differences, etc.).

In memory research, the issue of correlated memory signals has most often been considered in a two-alternative forced-choice (2-AFC) recognition task using similar targets and foils (Hall, 1979; Hintzman, 1988, 2001; Tulving, 1981). For example, the targets presented on the study list might consist of visual objects like an abacus, scissors, a picnic basket, and so on, and the corresponding foils might consist of a different abacus, different scissors, a different picnic basket, and so on. In one condition (correlated), the targets would be paired with their corresponding foils. In another condition (uncorrelated), the same targets and foils on the 2-AFC task would not be paired. Because the targets and foils are identical in both conditions, the distribution of memory signals they generate would be the same in both conditions. What would differ is the correlation between them. In the correlated condition only, participants can take advantage of the fact that the target, whether its memory-match signal is weak or strong, will typically generate a slightly stronger memory-match signal than its similar foil (Figure 6).

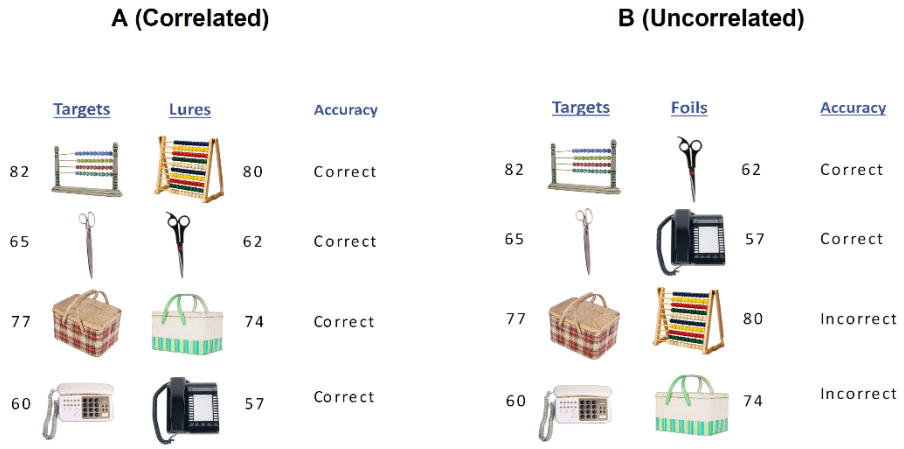


Figure 6. A (Correlated). Four 2-AFC test trials (top to bottom), with each target paired with its similar foil. The numbers to the left of each target and to the right of each foil represent the item's hypothetical memory-match signal. Note that each similar foil generates a memory-strength signal nearly as strong as its corresponding target because the two test items share many features. Although their strengths are similar,

the target in this idealized example can always be correctly chosen because it generates a slightly stronger memory signal than its similar foil. The target's signal is slightly stronger because only it provides a perfect match to memory. B (Uncorrelated). Everything is the same except that the targets are not paired with their similar foils. Now, the dissimilar foils sometimes generate a stronger memory signal than the target, leading to errors that would not occur in the correlated condition.

Although we have previously highlighted the importance of correlated memory signals for understanding lineup memory (Wixted et al., 2018, 2021), we have not modeled correlated signals in terms of feature matching. Yet a feature-matching mechanism is needed to understand why increasing filler similarity reduces empirical discriminability (illustrated earlier in Figure 2) even though it undoubtedly also increases the degree to which latent memory signals are correlated (which, as just noted, usually enhances discriminability). That seems paradoxical, but the paradox will be resolved once we specify the correlation in terms of the number of features that match across the faces in the lineup and then implement that result in signal detection models of eyewitness identification.

Assuming independence, the number of features with shared settings between any two faces in a lineup, which we represent as m, is equal to the number of features that match by design, n_D (i.e., the number of description-matched features), plus the number of remaining features, $n - n_D$, that match by chance. Because there are 5 settings for each feature, a chance match occurs with probability 1 / 5 = .20. Thus, $m = n_D + .20(n - n_D) = 5 + .20(20 - 5) = 8$. This equation applies regardless of the status of the two faces in the lineup (e.g., the guilty suspect and a filler, or the innocent suspect and a filler, or any two fillers). The guilty suspect has no special status for this measure because whether or not shared facial features also match memory is an independent consideration. Because m represents the number of features that match between any two faces in the lineup, its value applies equally to TA and TP lineups.

As noted earlier, we assume that features with shared settings across faces (e.g., blue eyes) generate the same feature-level memory-match signal. If the setting for shared feature f_i

happens to match memory of the perpetrator, then, for both faces, the feature-level memory-match signal would be drawn from the old distribution in Figure 3A (which has mean and standard deviation of 1). As an example, for a description-matched feature shared by everyone in a particular lineup (e.g., for f_1 = race/ethnicity, the setting for everyone might be Caucasian), the feature-level memory-match signal might be 1.46. In that case, the memory-match signal generated by this shared feature would be 1.46 for every face in the lineup.

A similar story applies if the setting for shared feature f_i happens *not* to match memory of the perpetrator. This could occur when the features of two faces in the lineup happen to match each other by chance but do not match the corresponding feature of the perpetrator's face in memory. For example, if f_9 = eye color, the setting for the perpetrator in memory might be P_9 = blue eyes, whereas the settings for both the innocent suspect (I) and a filler (F) might be $I_9 = F_9$ = green eyes. Thus, although the setting for f_9 is shared across faces, it does not match memory. The memory signal generated by this shared feature would be drawn from the new (foil) distribution in Figure 3A, which has a mean of 0 and a standard deviation of 1. For example, it might be -0.17, in which case the feature-level memory signal would be equal to -0.17 for each face in the lineup sharing that non-matching feature (i.e., green eyes).

In contrast to shared feature settings (which may or may not match memory), unshared feature settings between two faces in a lineup generate two different, independent memory signals (i.e., with independent error). If the feature of one face does not match memory (e.g., I_9 = green eyes), then its memory signal is drawn from the feature-level new distribution with a mean of 0 (Figure 3A). If the corresponding feature of the other face happens to match memory by chance (e.g., F_9 = blue eyes), then its memory signal is independently drawn from the feature-level target distribution with a mean of 1. The key point is that the two memory signals are

independent, regardless of which feature-level memory-strength distribution they are drawn from (new or old).

Each of the $n = 20$ features contributes variance of $\sigma_e^2 = 1$ to the summed memory signal of a given face such that $\sum_1^n \sigma_e^2 = \sigma^2 = n$ and, therefore, $\sigma = \sqrt{n}$ (as illustrated earlier in Figures 4 and 5). Critically, this is true whether or not the feature settings are shared with other faces in the lineup and whether or not they match memory. Regardless, the feature contributes $\sigma_e^2 = 1$ (on average) to the variance of the overall summed memory signal for the face. Therefore, some of the variance in the overall memory signal for a face comes from features with shared settings across the faces in the lineup and some comes from features with unshared settings. This feature-matching scenario is illustrated in Figure 7.

	TP lineup		TA lineup	
features	filler	guilty	filler	innocent
1	**1**		**1**	
2	**1**		**1**	
3	**1**		**1**	
4	**1**		**1**	
5	**1**		**1**	
6	**1**		0	
7	**1**		0	
8	**1**		0	
9	0	**1**	0	0
10	0	**1**	**1**	0
11	0	**1**	0	0
12	0	**1**	0	**1**
13	0	**1**	0	0
14	0	**1**	0	**1**
15	0	**1**	0	0
16	0	**1**	0	0
17	0	**1**	**1**	0
18	0	**1**	0	**1**
19	0	**1**	**1**	0
20	0	**1**	0	0
Σ	8.00	20.00	8.00	8.00

Figure 7. Each entry represents the mean of the distribution from which a feature-level memory-match signal is drawn (1 or 0, depending on whether the feature matches memory) for a two-person TP lineup and a two-person TA lineup. Features $f_1 \rightarrow f_5$ have settings that match each other by design because they were included in the witness's description. Because these shared features also match memory of the perpetrator, their

shared memory signals are drawn from the feature-level target distribution with a mean of 1 (Figure 3A). Of the remaining features ($f_6 \rightarrow f_{20}$), three feature settings ($f_6 \rightarrow f_8$) match each other by chance. In a TP lineup, these coincidentally matching features also match memory (so their shared memory signals are drawn from a distribution with a mean of 1), whereas in a TA lineup, the three coincidentally matching features match memory by chance (with probability $1/v$ = .2). In this example, $f_6 \rightarrow f_8$ for the TA lineup happen to not match memory. Therefore, their shared memory signals are drawn from the feature-level foil distribution with a mean of 0 (Figure 3A). Of the remaining features ($f_9 \rightarrow f_{20}$) in the TA lineup, three match memory of the perpetrator by chance. This is independently true of the filler and the innocent suspect. The average strength of the overall memory-match signal for a face in a lineup (shown at the bottom) is the sum of the 20 feature-match signals. Because 20 feature-match signals are summed for every face, the standard deviation of the memory signal is always $\sqrt{20}$.

For features that are shared across faces in a lineup ($f_1 \rightarrow f_8$ in Figure 7), all of the variance in the memory-match signals they generate occurs between lineups because these features have no within-lineup variance (i.e., within a lineup, the feature-level memory signal is the same). This means that part of the variance in the summed signal across the 20 features (i.e., part of σ^2) reflects between-lineup variance. We denote the between-lineup variance component of the summed memory signal σ_b^2. Memory signals generated by the remaining unshared features are independent, so the variance across those signals occurs within lineups. This means that the remaining part of σ^2 reflects within-lineup variance. We denote the within lineup variance component of the summed memory signal σ_w^2. Thus, $\sigma^2 = \sigma_b^2 + \sigma_w^2$. The relevance of this equation lies in the fact that the correlation between the overall (summed) memory signals generated by faces across lineups (ρ) is given by $\rho = \sigma_b^2 / (\sigma_b^2 + \sigma_w^2)$.

We went to the trouble of explaining all of this because it puts us in a position to quantify the correlation between face-level memory signals in terms of feature-matching. Because m features across two faces in a lineup share variance (whether or not they match memory), and because each of those features contributes $\sigma_e^2 = 1$, on average, the overall contribution of shared variance is $\sigma_b^2 = m = 8$. The remaining $n - m$ features do not have matching settings, so they contribute unshared variance such that $\sigma_w^2 = n - m = 12$. Thus, $\sigma^2 = \sigma_b^2 + \sigma_w^2 = 8 +$

$12 = 20$. Because $\rho = \sigma_b^2 / (\sigma_b^2 + \sigma_w^2)$, in terms of feature-matching, it follows that $\rho = m / (m + n - m) = m / n$. The intuitively appealing implication is that the correlation between the summed memory-match signals of the faces in a lineup is equal to the proportion of facial features with settings that match across the faces in a lineup. In our running example (where $m = 8$ and $n = 20$), the correlation between the summed memory signals generated by faces in a lineup would be 8 / (8 + 12) = 8 / 20 = .40. In the extreme high-similarity case where $m = n = 20$, the faces would be identical, in which case the correlation would equal 1.

Critically, experimentally manipulating filler similarity is conceptualized as manipulating *m* (for further illustration of these ideas, see Appendix: Correlated Memory Signals). Thus, writing the equations for competing signal detection models in terms of m and n, as we do next, makes it easy to see what those models predict about the effect of manipulating filler similarity on the degree to which the underlying memory signals associated with innocent from guilty suspects overlap (i.e., d'_{IG}). The equation we just worked out relating the correlation coefficient to the proportion of matching facial features (i.e., $\rho = m / n$) will constitute part of the model-specific equations we consider next.

Models of eyewitness decision-making

Wixted et al. (2018) derived the equations for three competing signal detection models of lineup performance: the Independent Observations model, the Ensemble model, and the Integration model. In this section, we present those previously derived equations, all of which include ρ as a parameter. We then translate the equations into feature-matching terms to specify what they predict about how manipulating filler similarity should affect d'_{IG}. This is the key prediction because the models unequivocally disagree about what should happen. Once those predictions are specified, we then fit the models (in their original forms) to empirical data to test

how manipulating filler similarity affects the degree to which the memory signals associated with innocent and guilty suspects overlap (as quantified by d'_{IG}).

Independent Observations model

The simplest decision variable is the raw memory-match signal generated by a face in the lineup (i.e., the face-level memory signal created by summing over feature-level memory-match signals) that we have been considering thus far. In a lineup, one of the faces will generate the strongest overall memory signal (the MAX face), and the Independent Observations model holds that if the memory signal of the MAX face in the lineup exceeds a decision criterion, that face will be identified, regardless of how strong the signals generated by the other faces happen to be (Macmillan & Creelman, 2005; Wixted et al., 2018). This decision variable was referred to as the Best-Above-Criterion Model by Clark et al. (2011). The stronger the memory signal generated by the MAX face is, the more confident the eyewitness will be when making an ID. In other words, according to this model, the raw memory-match signal is also the decision variable (just as is true of a showup).

Earlier, in Figure 5, we noted that:

$$d'_{TP} = \frac{\mu_G - \mu_{F_{TP}}}{\sigma} \quad (1)$$

However, this equation implicitly assumes uncorrelated memory signals, which must be true in a showup because at least two faces are needed in an identification procedure for memory signals to be correlated. However, it is unlikely to be true in a lineup. As noted by Wixted et al. (2018), *within* a lineup, the relevant standard deviation in the denominator is σ_w. As written, Equation 1 corresponds to the uncorrelated case because, in that case, $\sigma_w = \sigma$. More specifically, recall that $\sigma^2 = \sigma_b^2 + \sigma_w^2$ and that $\rho = \sigma_b^2 \,/\, (\sigma_b^2 + \sigma_w^2)$. In the uncorrelated case, $\sigma_b^2 = 0$ such that $\sigma^2 = \sigma_w^2$

and $\sigma_w = \sigma$. Thus, in the case of lineups, it would be better to write Equation 1 in its more general form:

$$d'_{TP} = \frac{\mu_G - \mu_{F_{TP}}}{\sigma_w} \quad (2)$$

Lineups presumably involve correlated memory signals such that $\sigma_b^2 > 0$. We know that $\sigma^2 = \sigma_b^2 + \sigma_w^2$, which means that $\sigma_b^2 = \sigma^2 - \sigma_w^2$. We also know that $\rho = \sigma_b^2 \,/\, (\sigma_b^2 + \sigma_w^2)$. Replacing σ_b^2 in this equation with $\sigma^2 - \sigma_w^2$ yields $\rho = (\sigma^2 - \sigma_w^2) \,/\, \sigma^2$. Solving this equation for σ_w^2 yields $\sigma_w^2 = \sigma^2 - \rho\sigma^2 = \sigma^2(1-\rho)$, or $\sigma_w = \sigma\sqrt{1-\rho}$. Finally, replacing σ_w in Equation 2 with $\sigma\sqrt{1-\rho}$ yields

$$d'_{TP} = \frac{\mu_G - \mu_{F_{TP}}}{\sigma\sqrt{(1-\rho)}} \quad (3)$$

Note that, as $\rho \to 1, d'_{TP} \to \infty$, which means that as the correlation increases, d'_{TP} increases (for reasons illustrated earlier in Figure 6). Equation 3 is based on the raw memory signals generated by the faces in the lineup (after taking into account the effect of the correlation) and therefore corresponds to the Independent Observations model. In terms of feature-matching, we noted earlier that μ_G and $\mu_{F_{TP}}$ (the mean memory-match signals for guilty suspects and fillers in a TP lineup, respectively) are equal to n_G and $n_{F_{TP}}$ (the corresponding number of feature settings that match the settings of the perpetrator in memory). Thus, we can substitute μ_G and $\mu_{F_{TP}}$ in Equation 3 with n_G and $n_{F_{TP}}$, respectively. For the terms in the denominator, we also know that $\sigma = \sqrt{n}$, and as noted in the previous section, in a description-matched lineup, $\rho = m \,/\, n$. Thus, expressed in terms of feature-matching, Equation 3 can be written as:

$$d'_{TP} = \frac{n_G - n_{F_{TP}}}{\sqrt{n}\sqrt{1 - m/n}}$$

In a TP lineup, n_G (number of guilty suspect feature settings that match memory) and n (number of features that define a face) are assumed to be equal, so we can replace n_G in this expression with n. In addition, because the guilty suspect is the perpetrator, the number of filler feature settings that match memory of the perpetrator ($n_{F_{TP}}$) is equal to the number of filler features that match the guilty suspect (i.e., the perpetrator) in the lineup (m). That is, for a TP lineup, $n_{F_{TP}} = m$. After making those substitutions in the numerator, we can write this equation as:

$$d'_{TP} = \frac{n - m}{\sqrt{n}\sqrt{1 - m/n}}$$

which reduces to the following simple equation:

$$d'_{TP} = \frac{n - m}{\sqrt{n - m}} \quad (4)$$

Equation 4 can be used to predict what should happen as filler similarity to the guilty suspect in a TP lineup is manipulated. From a pool of description-matched fillers, taking the additional step of selecting fillers who are also similar to the guilty suspect in a TP lineup (and, equivalently, to the perpetrator) is conceptualized as ensuring that more of the features $f_6 \rightarrow f_{20}$ have matching feature settings between the filler and the guilty suspect. Increasing filler similarity increases m (the number of filler features that match the guilty suspect), which simultaneously and equivalently increases $n_{F_{TP}}$ (the number of filler features that match memory of the perpetrator). This happens because increasing the number of features that match between a filler's face and the guilty suspect's face automatically increases the number of features that match between a filler's face and the face of the perpetrator in memory. Thus, $n_{F_{TP}} = m$.

Increasing m decreases the numerator of Equation 4, reflecting increasing similarity between a filler and the memory of the perpetrator, but is also decreases the denominator, reflecting the increased correlation between the memory signals generated by the faces in the

lineup. These two forces exert opposite effects on the degree to which memory signals associated with the guilty suspect and the fillers in a TP lineup overlap (i.e., d'_{TP}). However, the numerator decreases more rapidly than the denominator increases, which provides a theoretical explanation of why increasing filler similarity reduces discriminability within a TP lineup despite the increased correlation. By contrast, in Figure 6, which illustrated why correlated memory signals can enhance discriminability, only the correlation changed across conditions.

Critically, from a pool of description-matched fillers, taking the additional step of selecting fillers who are *dissimilar* to the guilty suspect in a TP lineup is conceptualized as reducing m (and, equivalently, reducing $n_{F_{TP}}$ and, therefore, $\mu_{F_{TP}}$), thereby increasing d'_{TP}. There would still be at least $n_D = 5$ matching feature settings because they were included in the description (i.e., the lineup would still be fair), but now fewer features would match by chance. In the extreme, m would be reduced from 8 to only 5 in our running example. According to Equation 4, this approach would maximize d'_{TP}. In short, to optimize discriminability in a TP lineup, from a pool of description-matched fillers, one should choose fillers that maximize dissimilarity to the suspect. Although Equation 4 is conceptually informative, we note that it can be further reduced to its simplest form as follows:

$$d'_{TP} = \sqrt{n - m} \tag{5}$$

Equation 5 unambiguously predicts that in a fair, description-matched TP lineup, the memory signals associated with the guilty suspect would be maximally separated from the memory signals associated with the fillers by minimizing m. In other words, maximizing dissimilarity would maximize the ability to discriminate the guilty suspect from the description-matched fillers in a TP lineup.

In a TA lineup, the degree to which the memory signals associated with the innocent suspect and the fillers overlap is given by:

$$d'_{TA} = \frac{\mu_I - \mu_{F_{TA}}}{\sigma\sqrt{(1-\rho)}} \quad (6)$$

Equation 6 is exactly analogous to Equation 3 for d'_{TP}. In a fair, description-matched TA lineup, before manipulating similarity, the innocent suspect is effectively just another filler. Thus, in terms of feature-matching, the mean memory-match signals of innocent suspects and fillers (μ_I and $\mu_{F_{TA}}$, respectively) are equal to the corresponding number of feature settings that match the settings of the perpetrator in memory (n_I and $n_{F_{TA}}$), where $n_I = n_{F_{TA}} = n_D + p(n - n_D) = 8$. Note that, in a TA lineup, $n_{F_{TA}}$(number of filler features that match memory) does not necessarily equal m (number of filler features that match the other faces in the lineup), so we cannot use m in place of $n_{F_{TA}}$, as we did for $n_{F_{TP}}$ in TP lineups. The same is true for n_I. The reason is that even if all 20 features match between a filler and the innocent suspect (such that $m = 20$, in which case it would be a lineup of clones), it would not change the number of features that also match memory of the perpetrator. Thus, we can replace μ_I with n_I and $\mu_{F_{TA}}$ with $n_{F_{TA}}$, but we cannot go further than that. Also, as before, in terms of feature matching, $\sigma = \sqrt{n} = \sqrt{20}$, and $\rho = m/n$. Making these substitutions yields:

$$d'_{TA} = \frac{n_I - n_{F_{TA}}}{\sqrt{n-m}}$$

Because $n_I = n_{F_{TA}}$ in a fair lineup, this equation reduces to $d'_{TA} = 0$, and this is true regardless of filler similarity (i.e., regardless of m). In other words, d'_{TA} should remain equal to 0 (i.e., the innocent suspect should not stand out regardless of filler similarity) when similarity is manipulated with respect to the suspect in a TA lineup.

Figure 8 illustrates the effect of manipulating filler similarity according to the Independent Observations model under the assumption that the memory signals are uncorrelated. Under those conditions, the use of low-similarity fillers should clearly enhance discriminability in TP lineups, but discriminability in TA lineups should remain unchanged (in accordance with the data shown earlier in Figures 1 and 2). In practice, however, the magnitude of the predicted effect in TP lineups would be reduced by the fact that the beneficial effect of correlated memory signals is lowest in the low-similarity condition, so d'_{TP} would not increase as much as would otherwise be the case (see Equation 3). Ironically, as we shall see, this model that so naturally predicts the observed empirical pattern in the unrealistic uncorrelated case struggles somewhat to explain it in the more realistic correlated case.

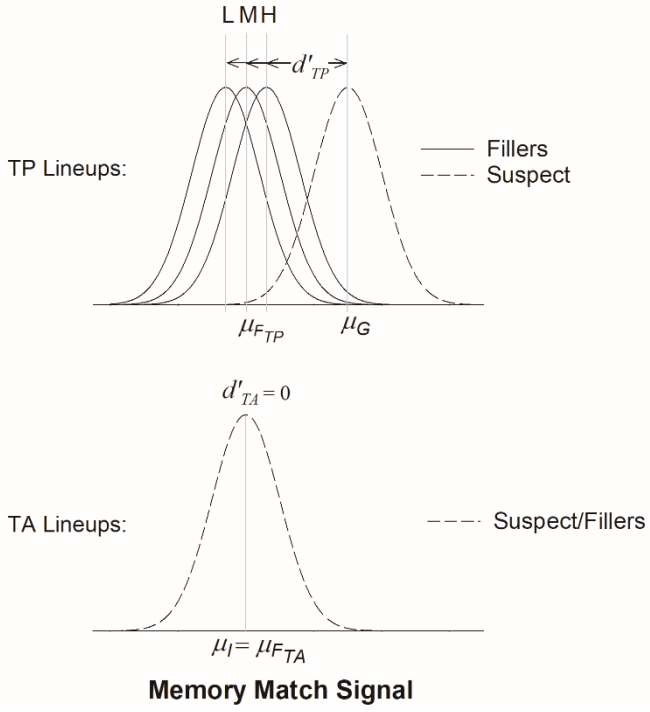


Figure 8. Distributions of memory signals in the low (L), medium (M), and high (H) filler-similarity conditions according to the Independent Observations model. These are the raw memory signals, before taking into account the effect of correlated signals.

We now turn to a key discriminability measure, which is the degree to which memory signals associated with innocent suspects in TA lineups overlap with memory signals associated with guilty suspects in TP lineups. Recall that this discriminability measure (illustrated earlier in Figures 4 and 5) is given by:

$$d'_{IG} = \frac{\mu_G - \mu_I}{\sigma}$$

This is the same equation that applies to a showup because, in the Independent Observations model, the memory signals generated by innocent and guilty suspects are independent of the presence or absence of fillers. The memory signals for innocent and guilty suspects are not correlated with each other because these faces appear in different lineups (i.e., there is no sense in which an innocent suspect's face from one lineup is paired with a guilty suspect's face from a different lineup for purposes of computing a correlation).

Replacing each term with the corresponding term of feature-matching yields:

$$d'_{IG} = \frac{n - n_I}{\sqrt{n}} \quad (7)$$

where n_I is the number of the innocent suspect's features that match memory. Note that m (the number of features of the suspect's face that match features of the fillers' faces) does not even appear in this equation. Filler similarity simply does not matter for this measure because the memory signals generated by the innocent and guilty suspects are *independent* of the memory signals generated by the fillers in their respective lineups.[4] Therefore, manipulating filler similarity (i.e., manipulating m) should not affect d'_{IG}.

[4] Note that, in this context, "independent" does not mean "uncorrelated." Instead, it means that the suspect memory signals in a TA or TP lineup are not affected by the memory signals generated by the fillers, whether the filler memory signals are correlated with the suspect memory signals or not.

Because d'_{IG} theoretically remains constant as a function of filler similarity according to the Independent Observations model, it might seem as though the model predicts no change in the ROC, which plots the hit rate (proportion of guilty suspect IDs from TP lineups) vs. the false alarm rate (proportion of innocent suspect IDs from TA lineups). However, if memory signals were uncorrelated, then (as illustrated earlier in Figure 8) this model unambiguously—and correctly—predicts that the empirical ROC (i.e., empirical discriminability) should increase with decreasing filler similarity (Colloff et al., 2021). It makes that correct prediction because of the effect on d'_{TP}, not d'_{IG}. Yet, as illustrated later, when we consider the effect of correlated memory signals, the predictions of this model on the empirical ROC as filler similarity is manipulated become more ambiguous.

Because the ROC data do not directly indicate the magnitude of d'_{IG}, the only way to determine if this latent measure of discriminability remains constant across filler-similarity conditions as predicted by the Independent Observations model is to fit the model to the relevant empirical data, such as to the data reported by Colloff et al. (2021). We do so in a later section after characterizing two competing signal detection models in terms of feature matching.

Ensemble model

The Ensemble model holds that the decision variable does not consist of the raw, untransformed memory signals generated by the faces in the lineup. Instead, this model assumes that the decision variable consists of the difference between the memory signals generated by a face and the average of the signals generated by all the faces in a lineup. As an example, for the singular face in the lineup that is a candidate for being identified (i.e., the MAX face), the relevant decision variable is the MAX signal minus the mean signal. If the MAX-minus-mean decision variable exceeds a decision criterion, the face will be identified, and the greater that

difference score is, the higher the confidence will be that the identified individual is the perpetrator.

This model sounds a bit more complicated than the Independent Observations model, but it also seems more plausible a priori. In the Ensemble model, a high-confidence ID will be made only if the memory signal of the MAX face stands out from the crowd of memory signals generated by the faces in the lineup. If every face in the lineup generates a strong memory signal, then the MAX face will not stand out, in which case an ID might not even be made (much less made with high confidence). In the Independent Observations model, by contrast, if the MAX face generates a strong memory signal, a high-confidence ID will be made even if the other faces also generate a strong memory signal. Intuitively, it seems unlikely that witnesses would rely on a decision variable that disregards the strength of the memory signals associated with the other faces in the lineup. Still, they might, and the Independent Observations model generally fits lineup data reasonably well (Wixted et al., 2018), which is why it remains a viable competitor.

Figure 9 shows the distribution of the relevant memory signals according to the Ensemble model for the uncorrelated case. Here, the mean memory signal in a lineup has been subtracted away from every raw memory-match signal, yielding transformed distributions.

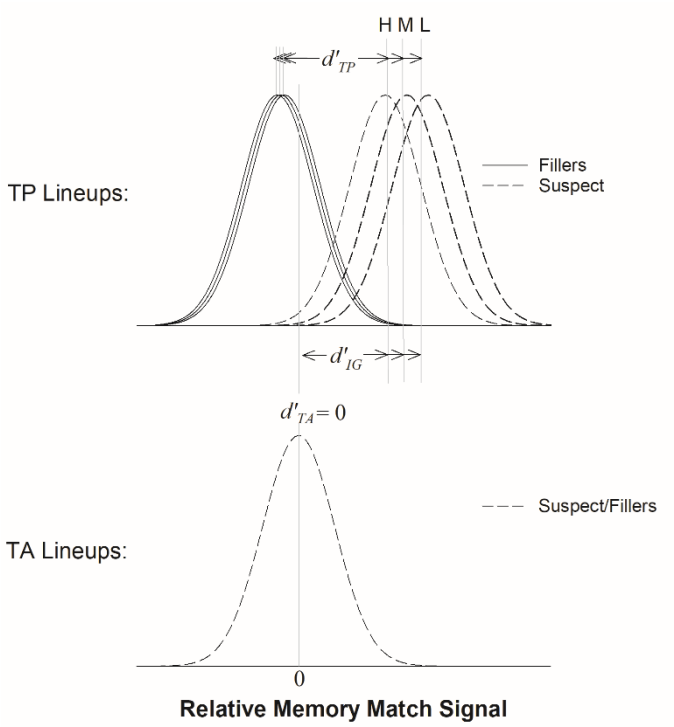


Figure 9. Operative memory signals when filler similarity is manipulated with respect to the suspect according to the Ensemble model in the low (L), medium (M), and high (H) filler-similarity conditions. This figure illustrates the distribution of transformed memory signals (each raw signal minus the mean signal for the lineup) without considering the effect of correlated memory signals.

As described in Wixted et al. (2018, 2021), the Ensemble model's equation for the ability to discriminate guilty suspect from fillers in a TP lineup is:

$$d'_{TP} = \frac{(\mu_G - \bar{\mu}_{TP}) - (\mu_{F_{TP}} - \bar{\mu}_{TP})}{\sigma\sqrt{(1-\rho)(1-1/k)}} \quad (8)$$

where $\bar{\mu}_{TP}$ is the mean memory signal across all members of the TP lineup (guilty suspect and fillers) and k is lineup size (i.e., k = 6 in a 6-person photo lineup). Because $\bar{\mu}_{TP}$ subtracts out of the numerator, Equation 8 reduces to:

$$d'_{TP} = \frac{\mu_G - \mu_{F_{TP}}}{\sigma\sqrt{(1-\rho)(1-1/k)}}$$

Replacing each term with its corresponding feature-matching term yields:

$$d'_{TP} = \frac{n_G - n_{F_{TP}}}{\sqrt{n}\sqrt{(1-m/n)(1-1/k)}}$$

Again, because $n_G = n$ and $n_{F_{TP}} = m$ in a TP lineup, we can make that substitution in the numerator. After doing so and multiplying $\sqrt{n}$ through in the denominator, we have:

$$d'_{TP} = \frac{n - m}{\sqrt{(n - m)(1 - 1/k)}}$$

Here again we see that reducing m (i.e., reducing filler similarity) has opposing effects in the numerator and denominator, but the ability to discriminate the guilty suspect from the fillers in the lineup should increase because the effect on the numerator is stronger. Simplifying further yields the final expression for d'_{TP} according to the Ensemble model:

$$d'_{TP} = \frac{\sqrt{n - m}}{\sqrt{1 - 1/k}} \quad (9)$$

Equation 9 is similar to the corresponding equation for the Independent Observations model (Equation 5) except it is divided by $\sqrt{1 - 1/k}$. Thus, in agreement with the intuitively sensible prediction made by the Independent Observations model, d'_{TP} should increase with decreasing filler similarity (i.e., with decreasing m).

For TA lineups, the equation for d'_{TA} according to the Ensemble model is directly analogous to the corresponding equation for d'_{TP} above:

$$d'_{TA} = \frac{(\mu_I - \bar{\mu}_{TA}) - (\mu_{F_{TA}} - \bar{\mu}_{TA})}{\sigma\sqrt{(1 - \rho)(1 - 1/k)}}$$

where $\bar{\mu}_{TA}$ is the mean memory signal across all members of the TA lineup (innocent suspect and fillers). This equation reduces to:

$$d'_{TA} = \frac{\mu_I - \mu_{F_{TA}}}{\sigma\sqrt{(1 - \rho)(1 - 1/k)}}$$

which, after replacing each term with its corresponding feature-matching term and simplifying, becomes:

$$d'_{TA} = \frac{n_I - n_{F_{TA}}}{\sqrt{(n-m)(1-1/k)}} \quad (10)$$

As before, $n_I = n_{F_{TA}}$ (i.e., the numerator is equal to 0), so $d'_{TA} = 0$. Moreover, neither n_I nor $n_{F_{TA}}$ is affected by manipulating filler similarity (i.e., by manipulating m). The reason is that in the ideal scenario we consider throughout, the innocent suspect is, in every important respect, just another filler. If $n_I = 8$ (i.e., 8 features of the innocent suspect match memory) and $n_{F_{TA}} = 8$ (i.e., 8 features of the innocent filler match memory), then 8 features will match memory no matter how many features they share between them. Therefore, d'_{TA} should remain equal to 0 regardless of m. This prediction from the Ensemble model matches the prediction made by the Independent Observations model.

Although the two models make the same basic predictions about d'_{TP} and d'_{TA} as a function of filler similarity, they diverge in what they predict about d'_{IG}. According to the Ensemble model (Wixted et al., 2018, 2021):

$$d'_{IG} = \frac{(\mu_G - \bar{\mu}_{TP}) - (\mu_I - \bar{\mu}_{TA})}{\sigma\sqrt{(1-\rho)(1-1/k)}}$$

where, again, $\bar{\mu}_{TP}$ and $\bar{\mu}_{TA}$ represent the average memory signals across all members of TA and TP lineups (respectively). This equation can be rearranged and simplified to:

$$d'_{IG} = (d'_{TP} - d'_{TA})(1-1/k) \quad (11)$$

Thus, unlike the Independent Observations model, where the value of d'_{IG} is not directly tethered to d'_{TP} and d'_{TA}, in the Ensemble model, an estimated value of d'_{IG} does not provide independent information beyond what is provided by d'_{TP} and d'_{TA}. To express Equation 11 in terms of feature-matching, we can replace d'_{TP} with the right side of Equation 9 and replace d'_{TA} with the right side of Equation 10 (which is equal to 0), yielding:

$$d'_{IG} = \frac{\sqrt{n-m}(1-1/k)}{\sqrt{1-1/k}}$$

which reduces to:

$$d'_{IG} = \sqrt{(n-m)(1-1/k)} \quad (12)$$

Thus, in contrast to the Independent Observations model, if the Ensemble model is correct, then d'_{IG} (like d'_{TP}) should increase with decreasing filler similarity (i.e., with decreasing *m*). In other words, according to this model, the degree to which the latent memory signals associated innocent and guilty suspects overall is minimized by maximizing filler dissimilarity.

Integration model

Like the Ensemble model, the Integration model holds that the decision variable does not consist of the raw, untransformed memory signals generated by the faces in the lineup. Instead, this model assumes that the decision variable consists of the *sum* of the memory signals generated by the faces in a lineup. If that summed decision variable exceeds a decision criterion, the MAX face will be identified, and the greater the magnitude of the summed decision variable, the higher the confidence will be that the identified individual is the perpetrator. A summed memory signal seems like an intuitively implausible decision variable (what sense does it make to confidently pick the MAX face because the faces all generate a strong memory signal?), but the Integration model has long been the only signal detection model that was used in the eyewitness identification literature (e.g., to compute d' from lineup data).

In contrast to the two models considered thus far, d'_{TP} and d'_{TA} are not expressible in terms of the Integration model. The reason is that the decision variable is equal to the sum of the memory signals over everyone in the lineup, so there is no within-lineup distinction between the memory signals for the suspect vs. the memory signals for the fillers. However, the summed

memory signals across everyone in the lineup still differ between TA and TP lineups, so d'_{IG} remains a relevant measure. As noted by Wixted et al. (2018), according to this model:

$$d'_{IG} = \frac{\left[\mu_G + (k-1)\mu_{F_{TP}}\right] - \left[\mu_I + (k-1)\mu_{F_{TA}}\right]}{\sigma\sqrt{k[1+(k-1)\rho]}} \tag{13}$$

After replacing each term in Equation 13 with the corresponding feature-matching term and simplifying, when filler similarity is manipulated with respect to the suspect, the equation becomes

$$d'_{IG} = \frac{\left[n_G + (k-1)n_{F_{TP}}\right] - \left[n_I + (k-1)n_{F_{TA}}\right]}{\sqrt{n}\sqrt{k[1+(k-1)m/n]}}$$

Once again, $n_{F_{TA}} = n_I$. After replacing $n_{F_{TA}}$ with n_I and rearranging, we have:

$$d'_{IG} = \frac{(n - kn_I) + (k-1)m}{\sqrt{kn + k(k-1)m}} \tag{14}$$

Wixted et al. (2018) found that this model fits the data poorly and can probably be rejected on those grounds alone. Here, we make the additional observation that, according to Equation 14, this model predicts the opposite of what the Ensemble model predicts about the effect of manipulating filler similarity on d'_{IG}. The Ensemble model predicts that increasing filler similarity (increasing m) will decrease d'_{IG}, whereas the Integration model predicts that it will increase d'_{IG} instead. To see why, note that the numerator of Equation 14 is of the form $\alpha + \beta m$ whereas the denominator is of the form $\sqrt{\delta + \lambda m}$, where the Greek letters are constants. According to Equation 14, d'_{IG} increases with increasing m because although m appears in both the numerator and the denominator (both of which increase with increasing m), the numerator increases faster than the denominator because the denominator is raised to a power of 0.5.

Model-based illustrations of manipulating filler similarity

To illustrate empirical ROC data anticipated by the three competing models, we first generated raw memory signals using the generic feature-matching model outlined earlier and then computed the appropriate decision variable separately for each model. For the Independent Observations model, the decision variable consisted of the raw memory signal itself, whereas for the Ensemble and Integration models, the raw signal was transformed in the manner described above to create the decision variable. The settings used to generate the raw memory signals were the setting used in our running example: $n_D = 5$, $n_G = n = 20$, $n_I = n_{F_{TA}} = n_D + .20(n - n_D) = 8$, and $n_{F_{TP}} = m$, where m = 5, 8, or 11 for the low-, medium, and high-similarity conditions, respectively. The degree to which the raw memory signals were correlated (m/n) is equal to $5/20 = .25$ in the low-similarity condition, $8/20 = .40$ in the medium-similarity condition, and $11/20 = .55$ in the high-similarity condition. The hypothetical ROC data generated by the three competing models are presented in Figure 10.

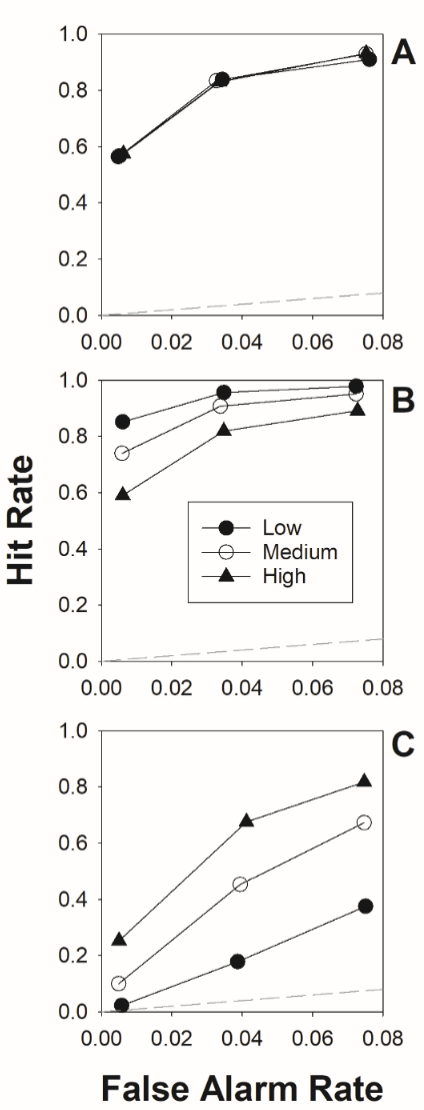


Figure 10. ROC data from the Independent Observations model (A), the Ensemble model (B), and the Integration model (C) based on raw memory signals generated by the feature-matching process, with parameters set to the values used in our running example.

The Independent Observations model (panel A), which correctly predicts that empirical *p*AUC will increase with decreasing filler similarity when memory signals are uncorrelated, surprisingly predicts no effect of filler similarity in this specific scenario (where memory signals become more correlated with increasing filler similarity). This change reflects a subtle effect arising from predicting empirical ROC data rather than just decision-variable separation

(captured by d'_{TA} and d'_{Tp}). Specifically, the probability that the suspect generates the maximum signal in the lineup and the probability that the maximum signal exceeds the decision criterion (thereby yielding an empirical hit or false alarm) are not entirely independent. As a result, even if d'_{TA} remains equal to 0 across filler similarity manipulations and even if the decision criterion remains constant, the false alarm rate will nevertheless vary to some degree as a function of how correlated the memory signals are. This subtle factor has less influence on the predictions made by the Ensemble model and Integration model (for both models, their respective d'_{IG} equations effectively guide thinking about the predicted empirical ROC), but it does have a more pronounced effect on the predictions made by the Independent Observations model, which show up as a discrepancy in predictions with (Figure 10A) and without (Figure 8) correlated memory signals.

Unlike the Independent Observations model, the Ensemble model (Figure 10B) predicts that discriminability will increase with decreasing similarity, in accordance with the predicted effect on d'_{IG} and in agreement with the empirical pattern depicted earlier in Figure 2. By contrast, the Integration model (Figure 10C) incorrectly predicts the opposite pattern (in accordance with its predicted effect on d'_{IG}). Thus, it seems that witnesses do not rely on the integration decision variable. We therefore once again conclude that this model is not viable.

Although the Independent Observations model also seems potentially challenged by the hypothetical results shown in Figure 10 (panel A), that pattern only demonstrates that the correct prediction it makes about manipulating filler similarity in the uncorrelated scenario would be eliminated if memory signals are differentially correlated in just the way envisioned by the specific parameters we used here (i.e., with $n = 20$, $n_D = 5$, and so on). The actual change in the correlation as a function of filler similarity is surely different in real lineups. Thus, to determine

which model is best able to account for empirical filler-similarity data, the models must be directly fit to such data.

Fitting the models to empirical data

We next fit the competing models to filler similarity data to test their predictions. As a reminder, these fits do not involve specific reference to the parameters of the feature-matching account (e.g., n, m, etc.), which were used to generate the predictions we test here and to provide a theoretical interpretation of each model's free parameters. Instead, the fits involve estimating the parameters of the models in their original form (e.g., estimating μ_G, $\mu_{F_{TP}}$, and $\mu_{F_{TA}}$, etc.). These parameters, once estimated, can be used to compute d'_{TP}, d'_{TA}, and d'_{IG} based on the equations presented earlier. The key question is how manipulating filler-similarity (i.e., manipulating m when conceptualized in terms of feature matching) affects the estimated value of d'_{IG}, with the relevant predictions specified by Equation 7 for the Independent Observations model and Equation 12 for the Ensemble model. According to those equations, the latent measure d'_{IG} should remain constant as a function of m if the Independent Observations model is correct, but it should increase as m decreases (high filler similarity to low filler similarity) if the Ensemble model is correct.

Data from Colloff et al. (2021)

To test the predictions worked out above, we first fit the Independent Observations model and the Ensemble model to the filler similarity data recently reported by Colloff et al. (2021), which were summarized earlier in Figures 1 and 2. We did not also fit the Integration model to those data because it predicts the opposite of the observed pattern, but we do provide a few tests of its predictions below (which are, characteristically, wildly off the mark).

Model parameters. For fits of both models, we defined μ_I to be 0 ($\mu_I \equiv 0$) and σ_I (standard deviation of the innocent suspect distribution) to be 1 ($\sigma_I \equiv 1$). We also set σ_G (standard deviation of the guilty suspect distribution) to 1 because allowing it to deviate from σ_I never significantly improved the fit. The Independent Observations model is characterized by seven additional parameters: μ_G (mean of the guilty suspect distribution), $\mu_{F_{TP}}$ (mean of the target-present filler distribution), $\mu_{F_{TA}}$ (mean of the target-absent filler distribution), σ_F (standard deviation of the filler distribution relative to the innocent and guilty distributions), and three confidence criteria, c_1, c_2, and c_3 (the confidence ratings were collapsed across adjacent ratings to create three confidence bins). Note that we would expect to find that $\sigma_F > 1$ because although the same innocent suspect was used in every TA lineup and the same guilty suspect was used in every TP lineup, the fillers in every similarity condition were randomly drawn from a pool of 109 faces. That random selection process would be expected to add random error to the filler distribution relative to the innocent and guilty suspect distributions (where $\sigma_G = \sigma_I \equiv 1$).

Across the three filler-similarity conditions, some of the seven parameters for the Independent Observations model were constrained to be equal, whereas others were free to vary. In particular, μ_G was constrained to be equal across similarity conditions (1 free parameter) because the guilty suspect was the same in every TP lineup, and, according to the foundational assumption of this model, the memory signals generated by the guilty suspect should be independent of the memory signals generated by the fillers. By contrast, we would expect $\mu_{F_{TP}}$ to vary systematically across the three conditions (Figure 5) because this is the parameter that was intentionally manipulated, so its estimated value was free to vary (3 parameters). We have no similar reason to expect $\mu_{F_{TA}}$ to vary across conditions or to deviate from $\mu_I \equiv 0$. Still, filler similarity to the innocent suspect was experimentally manipulated in TA lineups, so $\mu_{F_{TA}}$ might

(and, as it turns out, did) vary significantly as a result. We therefore did not constrain its value to be fixed across conditions (3 parameters). Similarly, σ_F was also initially free to vary across conditions (again because filler similarity was manipulated, perhaps affecting this parameter), but it never affected the fit significantly, so, while allowing its estimated value to deviate from 1, we constrained it to be equal across filler similarity conditions (1 free parameter). Finally, the three confidence criteria were allowed to vary across conditions (9 parameters). Thus, in all, there were 1 + 1 + 3 + 3 + 9 = 17 free parameters for this model. As described by Wixted et al. (2018), this model also has a correlation parameter, σ_b, but its estimated value never differed from 0 (even though it theoretically should have), so it plays no role in the fits of this model.

The parameters are similar for the Ensemble model but with a few key differences. The Ensemble model assumes that the operative psychological variable is a difference score, namely the difference between the face memory signal and the mean of the k memory signals generated by the faces in the lineup. This subtraction process untethers the values in TP lineups from the corresponding values in TA lineups (i.e., they are no longer measured with respect to $\mu_I \equiv 0$, as μ_G is in the Independent Observations model). When this model is fit to the data, the means of the raw memory signals (such as μ_G) are therefore not recoverable because the subtraction process has already theoretically taken place, before confidence ratings are provided.

In a TP lineup, the mean of the difference-score distribution for guilty suspects is $\mu_G - \bar{\mu}_{TP}$, where $\bar{\mu}_{TP}$ is the mean memory signal across all 6 faces in the lineup. Similarly, the mean of the difference-score distribution for the fillers is $\mu_{F_{TP}} - \bar{\mu}_{TP}$. The numerator of d'_{TP} for the Ensemble model (Equation 8) is $(\mu_G - \bar{\mu}_{TP}) - \left(\mu_{F_{TP}} - \bar{\mu}_{TP}\right)$, which reduces to $\mu_G - \mu_{F_{TP}}$. This difference score ($\mu_G - \mu_{F_{TP}}$), which is directly related to d'_{TP} in Figure 9, is all that can be estimated. For example, imagine that $\mu_G - \mu_{F_{TP}} = 1$. This could mean that $\mu_G = 2$ and $\mu_{F_{TP}} = 1$

or that $\mu_G = 3$ and $\mu_{F_{TP}} = 2$, but the numerator of Equation 8 would be the same either way. Thus, in fitting the model, all that matters is $\mu_G - \mu_{F_{TP}}$. This difference score is one free parameter, which we denote $\mu_{G-F_{TP}}$. Because the means of μ_G and $\mu_{F_{TP}}$ cannot be separately estimated, the Ensemble model has one fewer free parameter relative to the Independent Observations model. Similarly, for TA lineups, the only relevant memory-signal parameter for a given filler-similarity condition is $\mu_I - \mu_{F_{TA}}$, which we denote $\mu_{I-F_{TA}}$. However, a free parameter is not also lost here because μ_I is already fixed 0 by definition and is therefore not a free parameter.

The parameter $\mu_{G-F_{TP}}$ was free to vary across filler-similarity conditions (3 free parameters) and so was the parameter $\mu_{I-F_{TA}}$ (3 free parameters), whereas σ_F was again constrained to be equal across the three conditions (1 free parameter). The remaining parameters are the three confidence criteria (c_1, c_2, and c_3), and they were again free to vary across conditions (9 free parameters). Thus, the Ensemble model has 3 + 3 + 1 + 9 = 16 free parameters (one fewer than the basic Independent Observations model). This model does not have a correlation parameter because, theoretically, shared variance was subtracted away before the confidence rating was made. What remains is unshared variance, so the original correlated signals are theoretically unrecoverable.

Maximum likelihood fits. Table 1 shows the results of the maximum-likelihood fits. In terms of goodness of fit (χ^2, AIC, BIC), the two models were comparable. In addition, according to the Independent Observations model, the experimental manipulation of filler similarity in TP lineups was successful. That is, the estimated mean of the TP filler distribution ($\mu_{F_{TP}}$) became increasingly negative (i.e., the mean shifted more to the left, away from the guilty suspect

distribution) as filler similarity decreased from high to low. Constraining its value to be equal across the three conditions significantly worsened the fit, $\chi^2(2) = 91.1, p < .001$.

In TA lineups, the filler similarity manipulation also slightly affected $\mu_{F_{TA}}$ for reasons that are not clear. Constraining its value to be equal across the three conditions also significantly worsened the fit, $\chi^2(2) = 50.4, p < .001$. As expected, the estimated value of σ_F was significantly greater than 1, $\chi^2(1) = 122.8, p < .001$. The same trends were observed for the fit of the Ensemble model. That is, constraining $\mu_{G-F_{TP}}$ or $\mu_{I-F_{TA}}$ to be equal across conditions significantly worsened the fit, $\chi^2(2) = 86.6, p < .001$ and $\chi^2(2) = 7.84, p = .025$, respectively, as did constraining σ_F to equal 1, $\chi^2(1) = 96.8, p < .001$.

Table 1. Maximum likelihood parameter estimates and goodness-of-fit statistics for the Independent Observations model (A) and Ensemble model (B) fit to the filler-similarity data from the Suspect Similarity experiment reported by Colloff et al. (2021).

A

Model	Condition	μ_G	$\mu_{F_{TP}}$	$\mu_{F_{TA}}$	σ_F	c_1	c_2	c_3	N	npar	χ^2	AIC	BIC
Ind Obs	High		-0.18	0.00		1.59	2.14	2.87					
	Med	1.86	-0.52	-0.30	1.22	1.50	2.00	2.72	10559	17	55.0	30414.3	30537.8
	Low		-0.73	-0.28		1.46	1.93	2.65					

B

Model	Condition	---	$\mu_{G-F_{TP}}$	$\mu_{I-F_{TA}}$	σ_F	c_1	c_2	c_3	N	npar	χ^2	AIC	BIC
Ensemble	High		2.20	0.13		1.66	2.15	2.83					
	Med		2.57	0.36	1.37	1.83	2.28	2.97	10559	16	57.5	30415.1	30531.4
	Low		2.67	0.31		1.84	2.27	2.96					

We next computed the relevant discriminability measures from these parameter estimates (Table 2). These estimates are in line with predictions derived in the previous section in terms of feature matching. As a technical aside, for both models, some of these discriminability measures (namely, d_{TP} and d_{TA}) are not true d' scores because $\sigma \neq \sigma_F$ (for details, see Appendix: Computing Unequal-Variance Discriminability Measures). However, again for both models, d'_{IG} is a true d' score because only σ was involved in the calculation (where $\sigma = \sigma_G = \sigma_I = 1$). These technicalities aside, the discriminability measures for d_{TP} varied as a function of filler

similarity (i.e., as a function of m) in accordance with predictions, whereas the smaller and less systematic changes in d_{TA} were not anticipated.

Table 2. Discriminability statistics for the Independent Observations model (A) and Ensemble model (B) computed from the parameter estimates in Table 2.

A

Model	Condition	d_{TP}	d_{TA}	d'_{IG}
Ind Obs	High	1.83	0.00	
	Med	2.13	0.27	1.86
	Low	2.33	0.25	

B

Model	Condition	d_{TP}	d_{TA}	d'_{IG}
Ensemble	High	1.92	0.11	2.07
	Med	2.24	0.32	2.21
	Low	2.33	0.27	2.36

An unexpected outcome from the perspective of the Independent Observations model in Table 1 is that the estimates of $\mu_{F_{TP}}$ were negative in all three conditions. One would instead expect the estimate in the medium-similarity condition to approximately equal 0 (i.e., equivalent to μ_I) and to be positive in the high-similarity condition and negative in the low-similarity condition (Figure 8). Why would fillers who are high in similarity to the perpetrator generate a *weaker* memory-match signal, on average, than the innocent suspect (who, by design, has median similarity to the perpetrator)? Another unexpected outcome from the perspective of the Independent Observations model, noted earlier, is that the correlation parameter was always estimated to be 0 instead of increasing as a function of filler similarity.

Which model best accounts for the data? To explore a possible source of the unexpected parameter estimates provided by the Independent Observations model, we generated simulated data from the Ensemble model using its best-fitting parameter estimates in Table 1 and then fit the Independent Observations model to those Ensemble-generated data. The results are shown in the top three lines of data in Table 3, which now also includes a column for d'_{IG} (the estimate of

most interest). As above, $d'_{IG} = \mu_G$ because $d'_{IG} = \frac{\mu_G - \mu_I}{\sigma}$, and we have defined μ_I and σ_I to be 0 and 1, respectively (and $\sigma = \sigma_G = \sigma_I$). Interestingly, the estimates of $\mu_{F_{TP}}$ are now all negative, just as they are in the fit of the Independent Observations model to the real data.

Table 3. Maximum likelihood parameter estimates and goodness-of-fit statistics for the Independent Observations model fit to simulated data generated by the Ensemble model using the best-fitting parameter estimates for the Ensemble model presented in Table 1. Note that because $\mu_I = 0$ and $\sigma_I = \sigma_G = 1$, $d'_{IG} = \mu_G$

Model	Condition	d'_{IG}	μ_G	$\mu_{F_{TP}}$	$\mu_{F_{TA}}$	σ_F	c_1	c_2	c_3	N	npar	χ^2
Ind Obs	High	1.46	1.46	-0.20	0.03	1.18	1.48	1.99	2.67	20000	17	122.9
	Med			-0.47	-0.23		1.38	1.85	2.51			
	Low			-0.48	-0.26		1.35	1.80	2.46			
Ind Obs + 2	High	1.41	1.41	-0.22	0.02	1.16	1.43	1.93	2.57	20000	19	41.18
	Med	1.49	1.49	-0.44	-0.21		1.38	1.84	2.50			
	Low	1.64	1.64	-0.34	-0.12		1.47	1.92	2.58			

The bottom three lines of Table 3 show the results of allowing μ_G to vary across conditions, which means that d'_{IG} will vary as well. Doing so significantly improved the fit of the Independent Observations model, and the estimates of d'_{IG} were ordered in the direction predicted by the Ensemble model. This is perhaps not surprising given that the simulated data were actually generated by the Ensemble model (for which d'_{IG} necessarily varies as a function of filler similarity), but it raises a question: Would the same result be observed if μ_G were allowed to vary across conditions when the Independent Observations is fit the empirical data rather than to simulated data generated by the Ensemble model? If so, it would mean that d'_{IG} varies significantly across conditions even according to the Independent Observations model. Yet there is no reason why d'_{IG} should vary significantly or systematically if the Independent Observations model is correct.

As shown in Table 4, when μ_G was allowed to vary across conditions, the fit of the Independent Observations model to the empirical data was significantly improved, and the

estimates of d'_{IG} were ordered across conditions in the manner predicted by the Ensemble model (i.e., lowest in the high-similarity condition and highest in the low-similarity condition), $\chi^2(2) = 55.0 - 48.9 = 6.0, p = .047$. Note that this p value means that the d'_{IG} values differed significantly from each other without also taking into account that their values were ordered as predicted by the Ensemble model (an outcome that would occur by chance 1/6 of the time).

Table 4. Maximum likelihood parameter estimates and goodness-of-fit statistics for the Independent Observations model fit to the data from the Suspect Similarity experiment reported by Colloff et al. (2021). This time, μ_G was not constrained to be equal across conditions, which added two additional free parameters.

Model	Condition	d'_{IG}	μ_G	$\mu_{F_{TP}}$	$\mu_{F_{TA}}$	σ_F	c_1	c_2	c_3	N	npar	χ^2
Ind Obs + 2	High	1.79	1.79	-0.18	0.00		1.54	2.09	2.81			
	Med	1.89	1.89	-0.45	-0.24	1.20	1.52	2.02	2.73	10559	19	48.9
	Low	1.95	1.95	-0.61	-0.17		1.53	1.99	2.71			

Because only the Ensemble model predicts that d'_{IG} will be ordered this way, and because even the Independent Observations model interprets the data in accordance with prediction, the results would seem to provide compelling support for the Ensemble model. Then again, why would the two models fit the data about equally well if the Independent Observations model is incorrect (Table 1)? The answer appears to be that the Independent Observations model is considerably more flexible than the Ensemble model. That is, the Independent Observations model is better able to fit data generated by the Ensemble model than the other way around (see Appendix: Model Flexibility).

Detection ROCs. Another way to test the predictions of the competing models is to simply plot the "detection ROCs" from each filler-similarity condition. For this kind of ROC, a hit consists of any ID from a TP lineup (guilty suspect ID or filler ID) and a false alarm consists of any ID from a TA lineup (innocent suspect ID or filler ID). The three models make qualitatively different predictions about the order of the ROCs across the three filler-similarity conditions.

The Ensemble model predicts that the detection ROC curves will be ordered low-similarity > medium-similarity > high-similarity, as they were for the detection-plus-identification ROCs shown earlier in Figure 2. The prediction arises because in a TP lineup, fewer and fewer IDs should be made as the members of the lineup look more and more like the guilty suspect (because the difference-score decision variable becomes smaller, on average). The Independent Observations and Integration models both make the opposite prediction. That is, they both predict that the detection ROC curves will be ordered high-similarity > medium-similarity > low-similarity. According to the Independent Observations model, in a TP lineup, making the fillers more similar to the suspect should increase the chances that someone exceeds the decision criterion (elevating the hit rate). According to the Integration model, in a TP lineup, making the fillers more similar to the suspect should increase the summed decision variable, thereby also increasing the chances that someone exceeds the decision criterion (elevating the hit rate). No such effects should be observed in TA lineups.

Figure 11 shows the detection ROCs based on data reported by Colloff et al. (2021). Clearly, the high-similarity condition yields the *lowest* ROC, as uniquely predicted by the Ensemble model. The low-similarity condition very slightly exceeds the medium-similarity condition, but those two ROC curves basically fall atop one another. Although these results are not perfectly in accordance with predictions made by the Ensemble model, they are close. By contrast, the ROCs qualitatively differ from predictions made by the Independent Observations model and the Integration model.

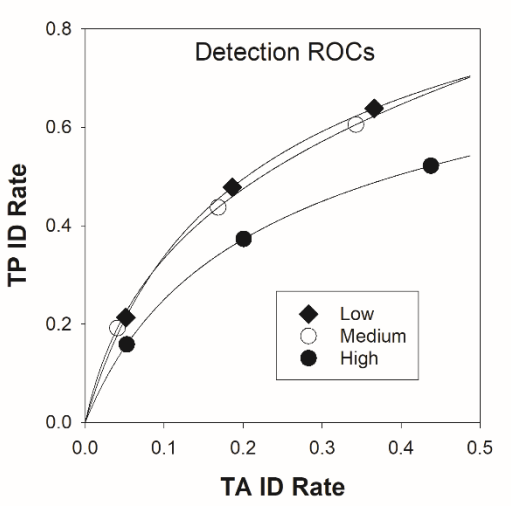


Figure 11. Detection ROCs from Colloff et al. (2021). For these ROCs, any ID from a TP lineup (to the guilty suspect or to a filler) was counted as a hit, whereas any ID from a TA lineup (to the innocent suspect or to a filler) was counted as a false alarm.

Overall, the results support the Ensemble model over the Independent Observations model (and over the already non-competitive Integration model). In more general terms, the results support the idea that eyewitnesses do not base their identification decisions on the memory-match signal generated by the MAX face in the lineup independent of the memory signals generated by the other faces in the lineup. Instead, these findings support the idea that witnesses take into account the other faces in the lineup.

Although the fits of the Independent Observations and Ensemble models to the data from Colloff et al. (2021) would appear to offer strong support for the Ensemble model, the results are based on only one specific stimulus set created to investigate the effect of manipulating filler similarity. To find out if the results generalize, we used a completely different approach to manipulate filler similarity in a new multi-trial-per-participant experiment reported here.

Data from a new filler-similarity experiment

The new experiment differed significantly from Colloff et al. (2021) in that filler similarity was manipulated by morphing the face of the suspect in the lineup onto the faces of the

fillers, to different degrees across filler similarity conditions. In addition, instead of being tested only once, each participant was tested multiple times (more like a traditional cognitive psychology experiment). Because the fillers were morphed to the suspect separately in TA and TP lineups, this experiment is analogous to the Suspect Similarity experiment reported by Colloff et al. (2021; i.e., the results presented earlier in Figures 1 and 2).

Participants. In total, 1,276 participants (M_{age}=34.12) were recruited through Amazon Mechanical Turk and included in the analysis for both successfully answering the attention check question and choosing “no” when asked “have you done this study before?”. The attention check question was “what were you asked to remember?” and the correct answer was “face”. The participants included 54.9% male (701), 44.3% female (565), 0.3% other (4) and 0.5% prefer not to state (6), with the ethnicity distribution being: 5% African-American (58), 14% Asian (184), 2% Mexican-American (22), 1% Filipino (15), 11% Latino (140), 3% Native-American (38), 56% Caucasian (709), 6% Other/Undeclared (78), and 3% Prefer not to state (32). The experiment was reviewed and approved by the University of California San Diego Social and Behavioral Sciences Institutional Review Board.

Materials and Design. All faces were selected from Chicago Face Database (CFD; Ma, Correll, & Wittenbrink, 2015). We used faces that were matched on the general characteristics commonly included in a witness’s description of a perpetrator (namely, race, age, gender, facial hair). In that sense, all of the fillers were considered to be description matched. The faces were Caucasian, male, approximately 30 years of age, and had no facial hair. Each face was cropped into an oval shape to exclude features such as hair and face shape because, to our eyes, including them made the morphed faces look morphed. Taking this approach could reduce real-world generalizability, but the key point of our study is to test theory-based predictions. Varying

methodological details that theoretically should not matter is arguably a good way to test the robustness of a model (e.g., Baribault et al., 2018).

The faces were randomly divided into six sets. Each set consisted of 15 faces, one of which was randomly picked and designated to serve as the suspect for that set (six suspects in all) and the other 14 of which served as potential fillers. We did not purposefully divide them based on similarity because they had to undergo the morphing process regardless.

The 14 non-suspect faces in each set were altered using Fantamorph software to create 3 pools of photos that varied in similarity to the suspect (low, medium, and high). These photos would serve as fillers in the lineups that contained the corresponding suspect. To create a pool of low-similarity fillers for a given set, the 14 non-suspect faces were morphed with the suspect to create new faces that were 20% suspect and 80% filler. To create a pool of medium-similarity fillers, the same 14 non-suspect photos were morphed with the suspect to create new faces that were 40% suspect and 60% filler. Finally, to create a pool of high-similarity fillers, the 14 non-suspect photos were morphed with the suspect to create new faces that were 60% suspect and 40% filler. In the end, we had six photo sets, with each set consisting of one suspect and 3 pools of low-, medium-, and high-similarity fillers (14 faces in each pool). Some examples are shown in Figure 12. Note that, in contrast to Colloff et al. (2021), these three conditions are actually best construed as three levels of high similarity because even in the “low”-similarity condition, the faces were morphed to be more similar to the suspect than they would be had the faces been unaltered. We refer to the three conditions of increasing filler similarity as 20% Suspect, 40% Suspect, and 60% Suspect. The model-based predictions about the effect of manipulating filler similarity remain unchanged.

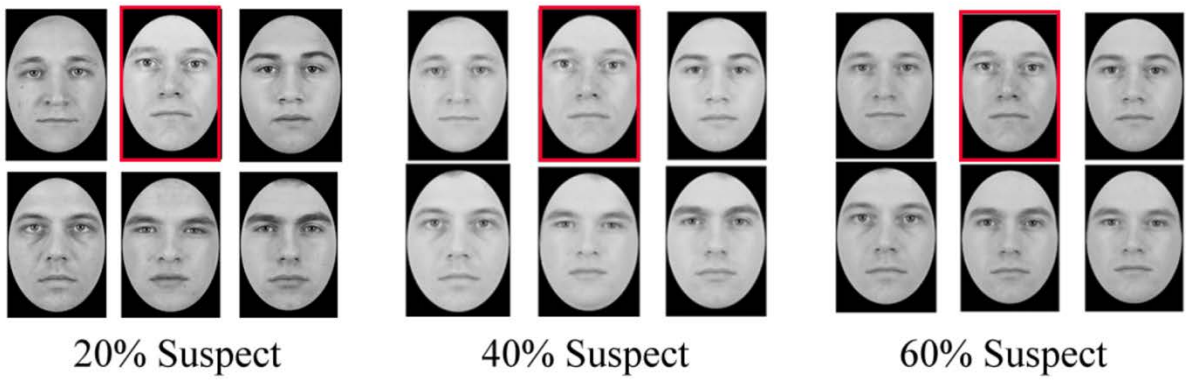
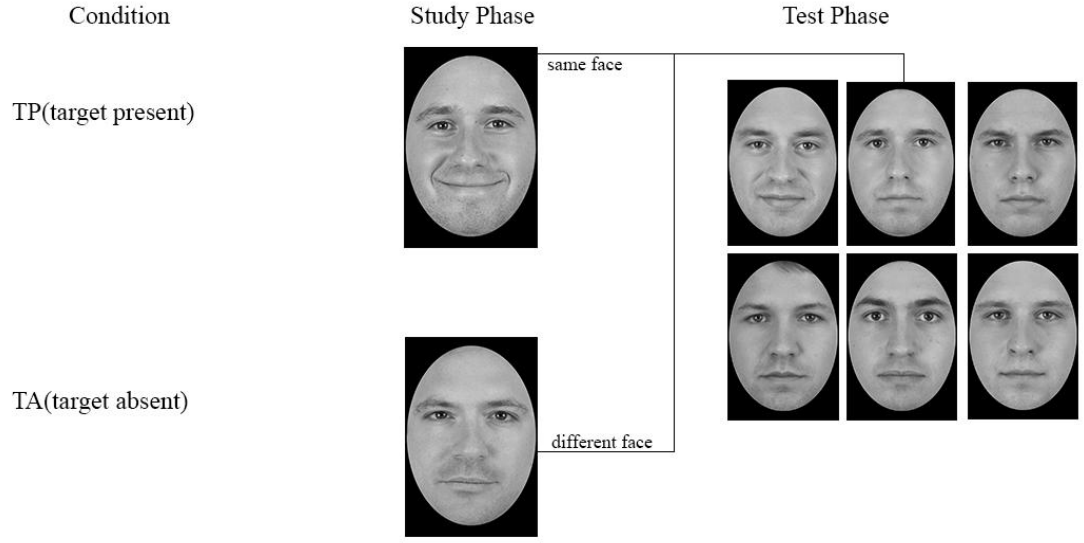


Figure 12. Examples of lineups constructed with stimuli at three similarity levels (morphed with 20%, 40%, 60% suspect), with the suspect being the middle face in the top row of each lineup. Note that, unlike Colloff et al. (2021), all three conditions involve a level of similarity above the medium level of similarity that would be obtained from choosing description-matched fillers.

One suspect photo from each of the six sets was used as the basis for the six lineups created for each participant. For a given participant, three of the six suspects were randomly assigned to be innocent (three TA lineups), and the remaining three suspects were assigned to be guilty (three TP lineups). For the three TA lineups, the suspects were randomly assigned to the low-, medium-, or high-similarity conditions, and the same was true of the suspects in the three TP lineups. For each lineup, five fillers were randomly drawn from the appropriate pool of 14 photos. As an example, if a suspect was assigned to the low-similarity TA condition, then five fillers were randomly drawn from the pool of low-similarity fillers that had been created for that suspect. Every participant received two lineups (one TP, one TA) for each of the three similarity levels in randomized orders. Both similarity and the presence of perpetrator were within-subject manipulations.

As illustrated in Figure 13, in the study phase of a TA lineup, the participant saw a photograph of a person who was randomly selected and not included in the six lineup photos for that trial, while in the study phase of a TP lineup, the participant saw a photograph of the suspect

that would appear in the six lineup photos for that trial. The study photo was not the exact same photo of the suspect in the TP lineup but was instead a photo of the same person with a different expression. Note that the overall design we used is a variant of the "single lineup paradigm" described by Oriet and Fitzgerald (2018) because the same lineup could be a TA lineup for one participant and a TP lineup for another. In other words, every lineup technically used seven faces total: the designated suspect, the five fillers that were associated with the suspect, and the perpetrator for the TA condition (when the designated suspect was innocent). If it was a TP lineup, the designated suspect was shown during the study phase. In that case, the lineup consisted of the designated suspect and five fillers. If it was a TA lineup, the perpetrator for the TA condition was shown during the study phase. In that case, the lineup consisted of the designated suspect and five fillers (the same lineup that they would have gotten had it been TP). Across participants, each suspect appeared in all possible conditions and no face was shown



more than once to the same subject, including the morphed fillers.

Figure 13. The procedure used in the experiment that manipulated filler similarity by morphing the face of the suspect onto the fillers to varying degrees.

Procedure. The study was programmed using Qualtrics and distributed through Amazon Mechanical Turk. The participants were first informed about the experimental procedure and asked to indicate consent if they would like to proceed. Then they received a practice trial before the six experimental trials began. The practice trial consisted of African American male suspects. Since the actual experiment used Caucasian male faces, no stimulus was shown to the subject more than once. Each trial consisted of three parts: study phase, distractor task and test phase. In the study phase, a photograph was shown for three seconds before the page auto-advanced. A mini game, either Tetris or a sliding block puzzle game called "2048," was then shown on the screen. The participant was instructed to score as high as possible. The game lasted for 60 seconds after which the test phase was presented. In the test phase, lineup members were presented simultaneously in a 2 x 3 array. The spatial location of each photograph was randomized. The program then instructed participants to either pick out the photograph of the person they previously saw or choose "none of the above". On the same screen, participants were asked to assess how confident they were about their identification decision using an 11-point scale, ranging from 0 (not certain at all) to 10 (absolutely certain). After all six trials concluded, participants were asked about their demographic information, what they were asked to study in the tasks (the attention check question), and whether they previously participated in this study.

Results. Table 5 presents the proportions of response outcomes (suspect ID, filler ID, or No ID) for TP and TA lineups across the three levels of filler similarity. In TP lineups, the suspect ID rate decreased substantially as similarity increased, while the filler ID rate increased and the No ID rate stayed consistent.

Table 5. Proportion of Guilty suspect IDs, Filler IDs, and lineup rejections in the 60% Suspect, 40% Suspect, and 20% Suspect conditions for Target-Present and Target-Absent Lineups

	TP			TA		
Condition	Guilty	Filler	Reject	Guilty	Filler	Reject
60%	0.25	0.22	0.53	0.06	0.26	0.68
40%	0.34	0.14	0.53	0.06	0.30	0.64
20%	0.39	0.08	0.53	0.08	0.20	0.72

A chi-square test was performed on each of the lineup categories, with filler IDs and No IDs collapsed due to our primary interest in suspect IDs. For the target-present lineups, there was a significant relationship between filler similarity and the number of suspect IDs, χ^2 (2, N = 3828) = 53.46, $p < .001$.[5] That is, the suspect ID rate increased as filler similarity decreased (from 60% to 20%), which is the same pattern recently reported by Colloff et al. (2021). For the target-absent lineups, there was no significant association between similarity and innocent suspect ID, χ^2 (2, N = 3828) = 3.89, $p = .14$. Again, this is the same pattern reported by Colloff et al. (2021). The results are consistent with our prediction that the ability to discriminate innocent from guilty suspects would increase as filler similarity decreased, and that pattern is evident in the ROC data shown in Figure 14.

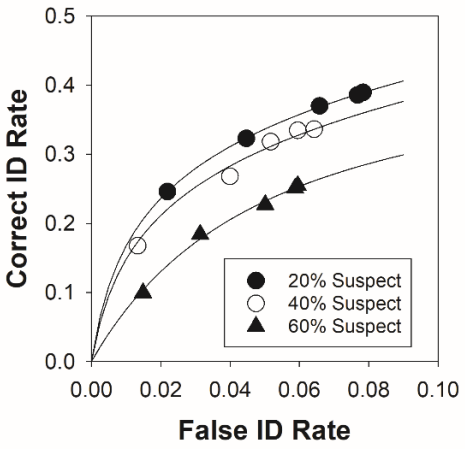


Figure 14. ROC data from the 20%, 40%, and 60% Suspect conditions of the face morphing study. Again, all three conditions involve fillers who are more similar to the suspect than the medium level of similarity that would be achieved by simply using description-matched fillers.

[5] The chi-square analyses for this experiment assume independence, which is not strictly true because each participant contributed 6 observations instead of the more typical 1.

Maximum likelihood fits. Using maximum likelihood estimation, we again fit the Independent Observations and Ensemble models to the frequency-count data for suspect IDs, filler IDs, and lineup rejections from TA and TP lineups. Note that, in this experiment, the fit was not improved by allowing the standard deviation of fillers and suspects to differ (i.e., an equal-variance model applied across the board), so one fewer parameter was involved in these fits compared to the ones reported earlier. This makes sense because a much smaller pool of filler faces was used in this experiment.

As shown in Table 6, the goodness-of-fit statistics were again comparable, and the parameter estimates showed the same basic patterns as before in Table 1. The parameter estimates for $u_{F_{TP}}$ in panel A confirm the effectiveness of the experimental manipulation, but their values are again hard to fathom from the perspective of the Independent Observations model. All of the values would be expected to be positive, with estimate in the low-similarity condition to be closest to 0 (i.e., closest to $u_I = 0$) because, in that condition, the fillers were only slightly morphed to the face of the guilty suspect. Nevertheless, the model appears to fit reasonably well.

Table 6. Maximum likelihood parameter estimates and goodness-of-fit statistics for the Independent Observations model (A) and Ensemble model (A) fit to the filler-similarity data.

A

Model	Condition	μ_G	μ_{F_TP}	μ_{F_TA}	c_1	c_2	c_3	npar	χ^2	AIC	BIC
Ind Obs	60%	1.07	0.01	0.03	1.58	1.93	2.34	16	51.5	19799.9	19911.0
	40%		-0.35	-0.05	1.42	1.71	2.06				
	20%		-0.65	-0.37	1.32	1.56	1.82				

B

Model	Condition	--	$\mu_G - \mu_{F\text{-}TP}$	$\mu_I - \mu_{F\text{-}TA}$	c_1	c_2	c_3	npar	χ^2	AIC	BIC
Ensemble	60%		1.01	0.06	1.40	1.70	2.04	15	58.8	19805.4	19909.6
	40%		1.30	0.03	1.42	1.66	1.97				
	20%		1.61	0.30	1.56	1.76	1.99				

The relevant discriminability measures computed from the parameter estimates in Table 6 (d_{TA}, d_{TP}, and d_{IG}) are shown in Table 7, and they show basically the same patterns as before in Table 2.

Table 7. Discriminability statistics for the Independent Observations and Ensemble models computed from the parameter estimates in Table 6.

Model	Condition	d'_{TP}	d'_{TA}	d'_{IG}
Ind Obs	60%	1.06	-0.03	
	40%	1.42	0.05	1.07
	20%	1.72	0.37	
Ensemble	60%	1.21	0.07	0.95
	40%	1.56	0.04	1.27
	20%	1.93	0.36	1.31

Which model was better supported by the data? Once again, as shown in Table 8, when we allowed μ_G for the Independent Observations model to vary across filler similarity conditions, the fit was significantly improved, $\chi^2(2) = 51.5 - 45.0 = 6.5, p = .039$. However, it is not just that μ_G differed significantly across the three filler-similarity conditions. Thus, because $\mu_G = d_{IG}$, the ability to discriminate innocent from guilty suspects varied systematically in the direction predicted by the Ensemble model. This once again means that even the fitted parameters of the Independent Observations model suggests that the Ensemble model is correct.

Table 8. Maximum likelihood parameter estimates and goodness-of-fit statistics for the Independent Observations model fit to the data. This time, μ_G was not constrained to be equal across conditions, which added two additional free parameters. Note that d'_{IG} is the same as μ_G for this equal-variance scenario.

Model	Condition	d'_{IG}	μ_G	μ_{F_TP}	μ_{F_TA}	c_1	c_2	c_3	npar	χ^2
Ind Obs + 2	60%	0.94	0.94	-0.07	-0.06	1.49	1.85	2.25		
	40%	1.08	1.08	-0.35	-0.06	1.42	1.71	2.06	18	45.0
	20%	1.17	1.17	-0.59	-0.30	1.38	1.62	1.89		

Detection ROCs. Finally, we also plotted the detection ROCs from this experiment (Figure 15). These ROCs plot the overall TP ID rate (guilty suspect and filler IDs combined) vs.

the TA ID rate (innocent suspect and filler IDs combined). The low-similarity condition again yields the highest ROC, as uniquely predicted by the Ensemble model, though the other two filler-similarity conditions fall essentially atop one another:

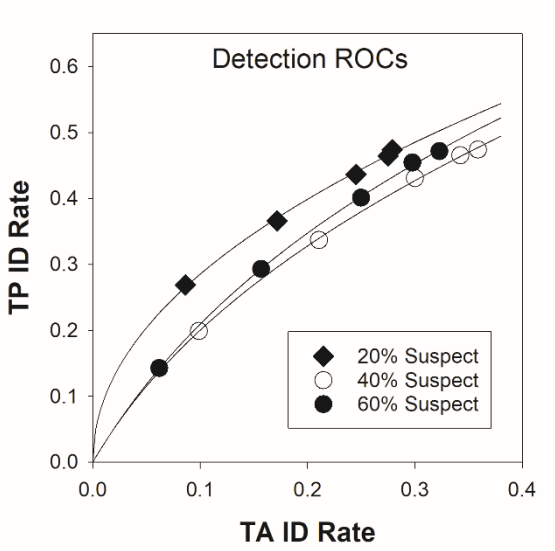


Figure 15. Detection ROCs from the face morphing study. Once again, for these ROCs, any ID from a TP lineup (to the guilty suspect or to a filler) was counted as a hit, whereas any ID from a TA lineup (to the innocent suspect or to a filler) was counted as a false alarm.

Results like these are difficult to reconcile with the Independent Observations model or the Integration model, both of which unambiguously predict that the low-similarity condition should yield the lowest (not the highest) detection ROC.

General Discussion

How do eyewitnesses deal with simultaneously generated memory signals to identify someone from the lineup or not? The answer to that question is related to another question that has bedeviled the field for decades: How similar should the fillers in a police lineup be to the suspect? Different signal detection models provide different answers to these questions. We tested competing models by expressing their predictions in terms of feature-matching and then fitting them to filler-similarity data. For the data sets considered here, the results suggest that the

decision variable is the degree to which the memory of one lineup member stands out from the memories of the other lineup members. The use of that decision variable predicts the non-obvious filler-similarity results observed by Colloff et al. (2021) and observed again here in a new study: from a pool of description-matched fillers, selecting those who are otherwise maximally dissimilar to the suspect maximizes ability to discriminability innocent from guilty suspects.

Optimizing filler similarity is a two-step process

Note that the first step of matching to the eyewitness's description of the perpetrator (before selecting dissimilar fillers) is critical. This step ensures that all faces in the lineup have features known to have encoded in the witness's brain. Skipping that step and maximizing dissimilarity per se would result in an unfair lineup. For example, if the perpetrator were described as a White male in his 30s, non-description-matched dissimilar fillers might consist of Hispanic men in their 20s, Black men in their 60s, and Asian men in their 40s. In that case, the suspect would differentially stand out in memory, whether innocent or guilty, harming the ability to discriminate innocent from guilty suspects (Colloff et al., 2016). Thus, an essential first step is to ensure that the pool of potential fillers matches the description of the perpetrator provided by the eyewitness, as has long been recommended (Wells et al., 1993, 2020).

Competing models of eyewitness identification

The simplest signal detection model of lineup memory is the Independent Observations model. Indeed, its simplicity is its most attractive feature. This model holds that witnesses base their decision (and their confidence) on the memory signal associated with the one face in the lineup that generates the strongest value (the MAX face). To date, the main rival to the Independent Observations model has been the Ensemble model, which holds that the decision

variable is instead the difference between the memory signal generated by the MAX face in the lineup and the mean memory signal generated by all the faces in the lineup. According to this model, the MAX face will be identified with high confidence not merely because it generates a strong memory signal but instead because it generates a *differentially* strong memory signal relative to the other faces in the lineup.

Both models can predict the pattern of results reported by Colloff et al. (2021), which is the same pattern observed here in a new experiment: when description-matched fillers are used across lineups, the ability of witnesses to distinguish between innocent and guilty suspects (i.e., empirical discriminability, measured by *p*AUC) increases the less similar the fillers are to the suspect. However, the two models differ in what they predict about the effect the filler-similarity manipulation on the distribution of underlying memory signals associated with innocent and guilty suspects (i.e., theoretical discriminability, quantified by d'_{IG}).

According to the Independent Observations model, the memory signals generated by suspects (innocent or guilty) are *independent* of the memory signals generated by the fillers in the lineup. Thus, manipulating the memory signals generated by the fillers should not affect d'_{IG}. Instead, the effect of filler similarity on empirical *p*AUC arises because reducing filler similarity increases d'_{TP} (the ability to discriminate the guilty suspect from the fillers in a TP lineup) while d'_{TA} (the ability to discriminate the innocent suspect from the fillers in a TA lineup) remains equal to 0. This effect is most clearly predicted when memory signals are uncorrelated (as illustrated in Figure 8). When memory signals are correlated across the faces in a lineup, as they presumably are, this model does not necessarily predict an effect of manipulating filler similarity (illustrated in Figure 10A).

According to the Ensemble model, by contrast, the operative memory signal generated by the suspect is *relative to* the mean of the memory signals generated by everyone in the lineup, including the fillers. Thus, manipulating filler similarity should affect d'_{IG}, which should be highest when low-similarity fillers are used and lowest when high-similarity fillers are used (Figures 9 and 10B). When the models were fit to the data, both agree that d'_{IG}, varied as a function of filler similarity in the manner predicted by the Ensemble model.

With regard to the Integration model, the results reported here would seem to effectively eliminate it from contention. This model was already challenged by its poor fit to data (Wixted et al., 2018), but it now suffers from the more fatal flaw of making directionally incorrect predictions about the effect of manipulating filler similarity with respect to the suspect. Yet most applications of signal detection theory to eyewitness identification over the years have relied on the Integration model (e.g., Duncan, 2006; Palmer & Brewer, 2012; Palmer, Brewer, Weber, & Sauer, 2020; Smith et al., 2018; Vitriol, Appleby, & Borgida, 2018). The model has generally been used to compute d', and it may often provide reasonable estimates. However, given that its predictions deviate so glaringly from empirical data, it would be better to compute underlying (i.e., theoretical) discriminability using the Independent Observations model or the Ensemble model.

Other findings bearing on the Ensemble model

The Ensemble model not only predicted the filler similarity findings considered here but is also consistent with a number of previously reported findings as well. A version of this model was originally proposed in verbal form to account for why simultaneous lineups yield higher discriminability than sequential lineups (Wixted & Mickes, 2014). According to “diagnostic feature-detection theory,” witnesses presented with a simultaneous lineup immediately realize

that certain features (namely, those included in the witness's description of the perpetrator and replicated across the lineup members) do not vary across the lineup members and are therefore non-diagnostic of guilt. Including those features in the decision variable would only add noise to the decision-making process without adding any signal, thereby decreasing the ability to discriminate innocent from guilty suspects. By instead discounting the common features (because they are of no help), error variance would be reduced, enhancing discriminability. In a sequential lineup, the faces are presented individually, so it is not as apparent to the eyewitness that non-diagnostic features were deliberately introduced when the photos were selected.

The Ensemble model can also explain why lineups are superior to showups. Because memory signals are correlated in lineups (but not showups), a subtractive decision variable can reduce noise relative to a showup, enhancing discriminability. In other words, compared to showups, the operative memory signals generated by guilty suspects in lineups overlap to a lesser degree with the memory signals generated by fillers/innocent suspects, thereby increasing d'_{TP} and d'_{IG}.

The Ensemble model also helps to explain why unfair lineups impair a witness's ability to discriminate innocent from guilty suspects compared to fair lineups (Colloff et al., 2016). In an unfair lineup, the suspect (innocent or guilty) matches the remembered features of the perpetrator better than the fillers do. These remembered features do not help to discriminate innocent from guilty suspects, yet if only the suspect has those features, they will be given weight by the eyewitness. Doing so can only add noise, increasing the degree to which the memory signals generated by innocent and guilty suspects overlap, reducing discriminability in that sense. When the remembered features of the perpetrator are replicated across everyone in a

description-matched lineup, by contrast, they will be discounted (i.e., shared variance will be subtracted out in terms of the Ensemble model), thereby enhancing discriminability.

Although these prior results are consistent with the Ensemble model, they do not rule out the use of the decision variable envisioned by the Independent Observations model. The reason is that those effects tend to also be predicted by the existence of correlated memory signals even if identification decisions are based on raw (i.e., untransformed) memory signals. For example, the fact that lineups reliably yield a higher area under the ROC (i.e., a higher pAUC) than showups does not contradict the Independent Observations model even though that model assumes that the memory signals generated by innocent and guilty suspects overlap to the same degree in lineups and showups. As noted earlier, correlated memory signals enhance d'_{TP}, and that consideration (i.e., correlated memory signals) can explain what the Independent Observations model does not explain in terms of overlapping memory signals. The filler-similarity findings reported here differ in that they directly contradict what the Independent Observations model predicts about d'_{IG}.

One previous finding that seems more consistent with the Independent Observations model than the Ensemble model involved manipulations of lineup size (k). The Ensemble model predicts that discriminability in TP lineups (d'_{TP}) will decrease with increasing lineup size, while the ability to discriminate innocent from guilty suspects (d'_{IG}) will increase with lineup size (Equation 12). Because both discriminability measures affect the empirical ROC, they might cancel, leaving the empirical ROC unaffected by lineup size. However, our simulations of the Ensemble model indicate that it usually predicts that the area under the empirical ROC should increase with lineup size, whereas the Independent Observations model makes no such prediction. Two recent studies found that although pAUC was greater for k = 2 (a 2-person

lineup) vs. $k = 1$ (a showup), pAUC did not significantly increase further for lineups of $k = 3$ up to $k = 12$ (Akan et al., 2020; Wooten et al., 2019). Both studies found a small trend in the direction predicted by the Ensemble model, but the results provide no compelling reason to reject the Independent Observations model. This result is not unlike findings from the ensemble coding literature where several studies found relatively constant sensitivity with increasing set size (Allik et al. 2013; Alvarez 2011; Ariely 2001; Chong & Treisman 2005).

The existing lineup size data are consistent with the Independent Observations model, which seems odd given that other evidence—including the new evidence presented here—that seems to favor the Ensemble model (Wixted et al., 2018). Indeed, multiple lineup studies have found that confidence in an ID is affected by the quality of the fillers, which suggests (but does not prove) that the decision may not be solely based on the MAX signal considered in isolation, as the Independent observations model assumes (Charman et al., 2011; Horry & Brewer, 2016). But if the Ensemble model is correct, why does a witness's ability to discriminate innocent from guilty suspects fail to increase beyond $k = 2$? As observed by Whitney and Leib (2018): "The benefit of averaging across larger sample sizes may be offset by factors such as increased correlated noise and positional uncertainty, potentially yielding a pattern of results that appears as if there is constant sensitivity across set sizes" (p. 115). And as noted by Mazyar, van den Berg, Seilheimer, and Ma (2013), Scottish philosopher Sir William Hamilton once observed that ''The greater the number of objects among which the attention of the mind is distributed, the feebler and less distinct will be its cognizance of each'' (Hamilton, 1859). In other words, the more items in the search set (here, the lineup), the larger σ will be, a factor not included in the equations for the Ensemble model. As σ increases, discriminability decreases. Thus, while the Ensemble model predicts an increase in the ability to discriminate innocent from guilty suspects

with increasing set size (with diminishing returns), Hamilton's law suggests that there may also be a countervailing force at play. Using a visual search task, Mazyar et al. (2013) found evidence suggesting that unless visual displays are largely predictable across trials (e.g., same distractors used over and over), the spreading of visual attention across items in the search set does indeed have detrimental effects on the quality of encoding of each stimulus. Perhaps something similar occurs as lineup size increases.

Alternatively, perhaps participants use a subtractive decision rule, as assumed by the Ensemble model, but they tend to focus on only two faces in the lineup, the MAX face and some other face (cf. Clark, 2011). If so, it would yield the discriminability benefits of subtracting shared variance but without predicting an increase in discriminability with increasing lineup size beyond $k = 2$. If this is true, then it would suggest that a method for further improving d'_{IG} would be to induce eyewitnesses to consider the MAX face in relation to all the faces in the lineup, not just to one other face.

Potential applied considerations

Finally, although our focus here is on theory, it is worth considering the potential applied implications of research on filler similarity.[6] Choosing dissimilar fillers from a pool of description-matched fillers enhances the ability to discriminate innocent from guilty suspects but, intuitively, may come across as being risky because it sounds like it should make the innocent suspect stand out. In fact, is easy to imagine cases where that might happen. For example, suppose the witness's description is: "clean-shaven White male in his early 30s with short brown hair." Every filler in the pool of description-matched photos would correspond to that

[6] We emphasize that we make no recommendations about actual police practices. Here, we have tested theoretical predictions of competing models under highly simplified conditions. Further research would be needed to assess the effects of maximizing filler dissimilarity to the suspect under the more varied conditions of the real world.

description, but some might have distinctive features, such as a conspicuous facial tattoo. In a target-absent lineup, if the dissimilar fillers were chosen because they had such distinctive features, then the innocent suspect would best correspond to the memory of the perpetrator (who presumably had no such feature given that it was not included in the description). It therefore might seem like choosing dissimilar fillers is a bad idea.

In both the lab and the real world, description-matched fillers who have distinctive features that were not described by the witness are typically removed from the pool of potential fillers, thereby minimizing this potential problem. However, this consideration leads to an interesting clarification of how the pool of potential fillers should be created. For decades, the rule has been to select *description-matched* fillers (i.e., fillers who match the description provided by the eyewitness). It seems that a more accurate rule—the one that might actually be used in practice—is to select fillers *who would have been described that way* (Frank & Goodman, 2012). Someone with a prominent distinctive facial feature like a tattoo likely would not have been described as "clean-shaven White male in high early 30s with short brown hair" and should therefore be excluded from the pool of potential fillers.

Another reason why the use of dissimilar fillers may sound risky is that, by chance, the innocent suspect might resemble the perpetrator more than the average potential filler does. Indeed, cases like that have been known to happen from time to time. An innocent suspect who happens to look a lot like the perpetrator would better match memory of the perpetrator than the fillers even if they were randomly drawn from the pool of potential fillers. Selecting fillers who are dissimilar to that innocent suspect will make that suspect stand out in memory even more, further increasing the chances of a false identification. To address this concern, does it therefore

make sense to do what the police sometimes do, which is to choose description-matched fillers who are also similar to the suspect?

We do not think so. Colloff et al. (2021) used the median-similarity filler from the pool of potential fillers as the innocent suspect. Half the time, an innocent suspect would more closely resemble the perpetrator (for them, using dissimilar fillers increases the risk of a false ID). The other half the time, an innocent suspect would less closely resemble the perpetrator (for them, using dissimilar fillers decreases the risk of a false ID). In the aggregate, the risk of a false ID should remain constant, as it has in the studies conducted thus far.

On the other hand, an innocent suspect's close resemblance to the perpetrator can arise for reasons other than random chance, such as when a suspect is selected based on a publicized photo or composite sketch. Taking this approach is a recipe for finding the perpetrator's innocent doppelganger. In that case, the innocent suspect would resemble the witness's memory of the perpetrator (more so than fillers selected based on the witness's description of the perpetrator). The use of dissimilar description-matched fillers under conditions like these—where a suspect is included in a lineup solely due to his or likeness to a publicized photo (with no evidence suggesting that this individual may have committed the crime)—would serve only to endanger innocent suspects (see Wells et al., 2020).

Then again, these considerations are not specific to maximizing filler dissimilarity. When the suspect was produced based on his or her resemblance to a publicized composite sketch or photo, eyewitness identification evidence will be biased against the suspect whether similar or dissimilar fillers are used. Thus, when someone becomes a suspect based on resemblance to a publicized image (with no independent evidence suggesting that the suspect may be the

perpetrator), it would make more sense to avoid using any kind of eyewitness identification procedure.

Conclusion

The research reported here is premised on the idea that cognitive models of lineup memory are essential to enhancing eyewitness accuracy now that decades of social psychology research has informed the field of what interpersonal factors can render police lineups ineffective and how to properly administer a lineup as a pure memory test devoid of social influence. Our research focuses on that particular scenario, regardless of how often this scenario applies in the real world, and even if it never applies in real world. When science-based recommendations are followed, a lineup becomes just another way of testing memory, like an old/new recognition test or a 2-alternative forced-choice test. The methodological details are slightly different, but the basic principles that have guided our theoretical understanding of those tasks apply nonetheless.

In our view, memory is memory, whether tested in the lab using a list of words or tested in the real world using a lineup. Treating a lineup as a test of memory, as we did here, resolved a question that has remained unanswered for decades: under ideal testing conditions, how similar should the fillers in a lineup be to the suspect and why? For good theoretical and empirical reasons, the two-step answer is as follows: (1) create a pool of fillers who would be reasonably described in the same way the witness described the perpetrator, and then (2) choose fillers from that pool who are maximally dissimilar to the suspect.

References

Akan, M., Robinson, M. M., Mickes, L., Wixted, J. T., & Benjamin, A. S. (2020). The effect of lineup size on eyewitness identification. *Journal of Experimental Psychology: Applied*. https://doi.org/10.1037/xap0000340

Allik, J., Toom, M., Raidvee, A., Averin, K., Kreegipuu, K. (2013). An almost general theory of mean size perception. *Vision Research, 83*, 25–39.

Alvarez, G. A. 2011. Representing multiple objects as an ensemble enhances visual cognition. *Trends Cognitive Sciences,15*, 122–31.

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

Baribault, B., Donkin, C., Little, D. R., Trueblood, J. S., Oravecz, Z., van Ravenzwaaij, D., White, C. N., De Boeck, P., & Vandekerckhove, J. (2018). Metastudies for robust tests of theory. *Proceedings of the National Academy of Sciences of the United States of America, 115*, 2607–2612.

Carlson, C. A., Jones, A. R., Whittington, J. E., Lockamyeir, R. F., Carlson, M. A., & Wooten, A. R. (2019). Lineup fairness: propitious heterogeneity and the diagnostic feature-detection hypothesis. *Cognitive Research: Principles and Implications, 4(1)*, 2.

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(6),* 629-654.

Clark, S. E. (2008). The importance (necessity) of computational modelling for eyewitness identification research. Applied Cognitive Psychology, 22(6), 803–813.

Clark, S. E., Erickson, M. A., & Breneman, J. (2011). Probative value of absolute and relative judgments in eyewitness identification. *Law and Human Behavior, 35*(5), 364–380.

Cohen, A. L., Starns, J. J., Rotello, C. M., & Cataldo, A. M. (2020). Estimating the proportion of guilty suspects and posterior probability of guilt in lineups using signal-detection models. *Cognitive Research: Principles and Implications, 5,* Article 21. https://doi.org/10.1186/s41235-020-00219-4

Colloff, M. F., Wade, K. A., & Strange, D. (2016). Unfair Lineups Make Witnesses More Likely to Confuse Innocent and Guilty Suspects. *Psychological Science, 27(9),* 1227–1239.

Colloff, M. F. (2021, February 16). Optimizing the selection of fillers in police lineups: Experiment 2. Retrieved from osf.io/c36bf

Colloff, M. F., Wilson, B. M., Seale-Carlisle, T. M., & Wixted, J. T. (2021). Optimizing the selection of fillers in police lineups. *Proceedings of the National Academy of Sciences,* 118(8).

Colloff, M. F. & Wixted, J. T. (2020). Why are lineups better than showups? A test of the filler siphoning and enhanced discriminability accounts. *Journal of Experimental Psychology: Applied, 26,* 124-143.

Cox, G. E., & Shiffrin, R. M. (2017). A dynamic approach to recognition memory. *Psychological Review, 124*(6), 795–860.

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

Dunn, J. C., Kaesler, M., & Semler, C. (2022). A model of position effects in the sequential lineup. *Journal of Memory and Language, 122,* 104297.

Fitzgerald, R. J., Oriet, C., & Price, H. L. (2015). Suspect filler similarity in eyewitness lineups: A literature review and a novel methodology. *Law and Human Behavior, 39(1),* 62-74.

Fitzgerald, R. J., Price, H. L., & Valentine, T. (2018). Eyewitness identification: Live, photo, and video lineups. *Psychology, Public Policy, and Law, 24(3),* 307–325.

Frank, M. C., & Goodman, N. D. (2012). Predicting Pragmatic Reasoning in Language Games. *Science, 336,* 998.

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review, 91(1),* 1–67.

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

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

Hamilton, W. (1859). Lectures on metaphysics and logic (vol. 1). Boston: Gould and Lincoln.

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

Kellen, D. & McAdoo, R. M. (in press). Towards a More Comprehensive Modeling of Sequential Lineups. *Cognitive Research: Principles and Implications*.

Kovera, M. B., & Evelo, A. J. (2021). Eyewitness identification in its social context. *Journal of Applied Research in Memory and Cognition, 10,* 313-327.

Lacroix, J. P., Murre, J. M., Postma, E. O., & Herik, H. J. (2006). Modeling recognition memory using the similarity structure of natural input. *Cognitive Science*, *30*(1), 121–145.

Lee, J., & Penrod, S. D. (2019). New signal detection theory-based framework for eyewitness performance in lineups. *Law and Human Behavior*, *43*(5), 436–454.

Lindsay, R. C. L., & Wells, G. L. (1980). What price justice? Exploring the relationship of lineup fairness to identification accuracy. *Law and Human Behavior*, *4,* 303–313.

Luus, C. A. E., & Wells, G. L. (1991). Eyewitness identification and the selection of distracters for lineups. *Law and Human Behavior*, *15*(1), 43–57.

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

Macmillan, N. A., & Creelman, C. D. (2005). *(2nd ed.). Lawrence Erlbaum Associates Publishers.*

Mazyar, H., van den Berg, R., Seilheimer, R. L., & Ma, W. J. (2013). Independence is elusive: Set size effects on encoding precision in visual search. *Journal of Vision*, *13*(5), Article 8.

Meltzer, M. A., & Bartlett, J. C. (2019). Holistic processing and unitization in face recognition memory. *Journal of Experimental Psychology: General, 148*(8), 1386–1406.

National Research Council. (2014). *Identifying the culprit: Assessing eyewitness identification.* Washington, DC: The National Academies Press. Retrieved from: https://www.nap.edu/catalog/18891/identifyingthe-culprit-assessing-eyewitness-identification

Nelson, A. B., & Shiffrin, R. M. (2013). The co-evolution of knowledge and event memory. *Psychological Review, 120(2),* 356–394.

Neuschatz, J. S., Wetmore, S. A., Key, K. N., Cash, D. K., Gronlund, S. D., & Goodsell, C. A. (2016). A comprehensive evaluation of showups. In B. Bornstein & M. K. Miller (Eds.), *Advances in psychology and law* (pp. 43–69). Cham, Switzerland: Springer International Publishing.

Nosofsky, R. M. (1991). Tests of an exemplar model for relating perceptual classification and recognition memory. *Journal of Experimental Psychology: Human Perception and Performance, 17(1),* 3–27.

Oriet, C., & Fitzgerald, R. J. (2018). The single lineup paradigm: A new way to manipulate target presence in eyewitness identification experiments. *Law and Human Behavior,* 42(1), 1-12.

Osth, A. F., & Dennis, S. (2015). Sources of interference in item and associative recognition memory. *Psychological Review, 122*(2), 260–311.

Palmer, M. A., & Brewer, N. (2012). Sequential lineup presentation promotes less-biased criterion setting but does not improve discriminability. *Law and Human Behavior, 36*(3), 247–255.

Palmer, M. A., Brewer, N., Weber, N., & Sauer, J. D. (2020). Eyewitness identifications of multiple culprits: Disconfirming feedback following one lineup decision impairs identification of another culprit. *Psychology, Public Policy, and Law.* Advance online publication.

Police Executive Research Forum. (2013). A national survey of eyewitness identification procedures in law enforcement agencies. Retrieved from: http://www.policeforum.org/assets/docs/Free_Online_Documents/Eyewitness_Identification/a national survey of eyewitness identification procedures in law enforcement agencies 2013.pdf

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J. C., & Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC bioinformatics, 12(1), 77.

Rotello, C. M., & Chen, T. (2016). ROC curve analyses of eyewitness identification decisions: An analysis of the recent debate. *Cognitive Research: Principles and Implications, 1(1),* 10. https://doi.org/10.1186/s41235-016-0006-7

Sauer, J. D., Brewer, N., & Weber, N. (2008). Multiple confidence estimates as indices of eyewitness memory. *Journal of Experimental Psychology: General, 137(3),* 528-547.

Shen, K. J. (2022, February 8). Modeling Face Similarity in Police Lineups. Retrieved from osf.io/fr4xd.

Shepard, R. N. (1958). Stimulus and response generalization: Tests of a model relating generalization to distance in psychological space. *Journal of Experimental Psychology, 55,* 509-523.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science, 237*, 1317-1323.

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM – retrieving effectively from memory. *Psychonomic Bulletin & Review, 4(2)*, 145–166.

Smith, A. M., Wells, G. L., Smalarz, L., & Lampinen, J. M. (2018). Increasing the similarity of lineup fillers to the suspect improves the applied value of lineups without improving memory performance: Commentary on Colloff, Wade, and Strange (2016). *Psychological Science, 29*(9), 1548–1551.

Starns, J. J., Cohen, A. L., & Rotello, C. M. (2021). A complete method for assessing the effectiveness of eyewitness identification procedures: Expected information gain. *Psychological Review.* Advance online publication. https://doi.org/10.1037/rev0000332

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

Vitriol, J. A., Appleby, J., & Borgida, E. (2018). Racial bias increases false identification of black suspects in simultaneous lineups. *Social Psychological and Personality Science, 10*, 722-734.

Wells, G. L., Kovera, M. B., Douglass, A. B., Brewer, N., Meissner, C. A., & Wixted, J. T. (2020). Policy and procedure recommendations for the collection and preservation of eyewitness identification evidence. *Law and Human Behavior, 44*, 3-36.

Wells, G. L., Rydell, S. M., & Seelau, E. P. (1993). The selection of distractors for eyewitness lineups. Journal of Applied Psychology, 78(5), 835.

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 & Cognition, 4*, 8–14.

Whitney, D., & Leib, A. Y. (2018). Ensemble perception. *Annual Review of Psychology, 69*, 105–129.

Wixted, J. T. (2020). The forgotten history of signal detection theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 46*, 201-233.

Wixted, J. T., & Mickes, L. (2012). The field of eyewitness memory should abandon probative value and embrace receiver operating characteristic analysis. Perspectives on Psychological Science, 7(3), 275-278.

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. (2015). ROC analysis measures objective discriminability for any eyewitness identification procedure. Journal of Applied Research in Memory and Cognition, 4(4), 329-334.

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

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

Wixted, J. T., Vul, E., Mickes, L. & Wilson, B. W. (2021). Eyewitness Identification is a visual search task. *Annual Review of Vision Science, 7*, 7.1-7.23.

Wooten, A. R., Carlson, C. A., Lockamyeir, R. F., Carlson, M. A., Jones, A. R., Dias, J. L., & Hemby, J. A. (2020). The number of fillers may not matter as long as they all match the description: The effect of simultaneous lineup size on eyewitness identification. *Applied Cognitive Psychology, 34*, 1-15.

Appendix

Correlated Memory Signals

In terms of feature matching, the degree to which the overall memory signals generated by faces in a lineup are correlated depends on the number of features shared across the faces in a lineup (whether or not those features also match memory of the perpetrator). Here, we further illustrate the basic idea in three figures (Figures A1, A2, and A3).

	TP lineup		TA lineup	
features	filler	guilty	filler	innocent
1	**1**		**1**	
2	**1**		**1**	
3	**1**		**1**	
4	**1**		**1**	
5	**1**		**1**	
6	**1**		0	
7	**1**		0	
8	**1**		0	
9	0	**1**	0	0
10	0	**1**	**1**	0
11	0	**1**	0	0
12	0	**1**	0	**1**
13	0	**1**	0	0
14	0	**1**	0	**1**
15	0	**1**	0	0
16	0	**1**	0	0
17	0	**1**	**1**	0
18	0	**1**	0	**1**
19	0	**1**	**1**	0
20	0	**1**	0	0
Σ	8.00	20.00	8.00	8.00

Figure A1. Each entry represents the mean of the distribution from which a feature-level memory-match signal is drawn (1 or 0, depending on whether the feature matches memory) for a two-person TP lineup and a two-person TA lineup. Features $f_1 \rightarrow f_5$ have settings that match each other by design because they were included in the witness's description. Because these shared features also match memory of the perpetrator, their memory signals are drawn from a distribution with a mean of 1. Of the remaining features ($f_6 \rightarrow f_{20}$), three features ($f_6 \rightarrow f_8$) match each other by chance. In a TP lineup, these coincidentally matching features also match memory (so their memory signals are drawn from a distribution with a mean of 1), whereas in a TA lineup, the three coincidentally matching features match memory by chance (with probability $1/v$ = .2). In this example, $f_6 \rightarrow f_8$ for the TA lineup happen to not match memory. Therefore, their memory signals are drawn from a distribution with a mean of 0. Of the remaining features ($f_9 \rightarrow f_{20}$) in the TA lineup, three match memory of the perpetrator by chance. This is independently true of the filler and the innocent suspect. The average strength of the overall memory-match signal for a face in a lineup (shown at the bottom) is the sum of the 20 feature-match signals.

	TP lineup		TA lineup	
features	filler	guilty	filler	innocent
1	**0.37**		**1.23**	
2	**1.44**		**0.37**	
3	**0.56**		**1.32**	
4	**-0.68**		**-0.01**	
5	**2.96**		**1.40**	
6	**2.66**		-0.47	
7	**0.64**		0.01	
8	**0.37**		0.16	
9	1.90	**-0.02**	0.29	-0.02
10	-0.66	**0.46**	**1.84**	-0.43
11	-0.13	**2.91**	-0.29	-0.51
12	0.13	**1.83**	0.01	**2.16**
13	0.55	**1.00**	1.11	1.03
14	-0.10	**0.54**	0.53	**1.99**
15	-1.20	**1.01**	-0.68	-1.13
16	-0.68	**0.71**	0.21	2.10
17	-1.02	**-0.01**	**1.56**	-0.54
18	-0.03	**1.70**	-0.54	**1.32**
19	0.88	**0.49**	**0.82**	-1.73
20	-0.67	**0.70**	-0.64	0.25
Σ	7.30	19.64	8.24	8.49

Figure A2. Hypothetical feature-level memory signals (now with random error) for a two-person TP lineup and a two-person TA lineup. Critically, shared features are assumed to share the random-variable memory signal they generate, which is the source of correlated summed memory signals. Values in bold represent feature settings that match memory and so are drawn from a distribution with a mean of 1, whereas non-bold values do not match memory and are drawn from a distribution with a mean of 0. It is the two summed signals at the bottom for a given lineup (e.g., 7.30 and 19.64 for the TP lineups) that are correlated across multiple lineups of that type. Correlated memory signals themselves are not illustrated here because this figure illustrates one TP lineup and one TA lineup. The summed signals are correlated (i.e., both high or both low) across lineups.

	Low				Medium				High			
	TP lineup		TA lineup		TP lineup		TA lineup		TP lineup		TA lineup	
features	filler	guilty	filler	innocent	filler	guilty	filler	innocent	filler	guilty	filler	innocent
1	**1**		**1**		**1**		**1**		**1**		**1**	
2	**1**		**1**		**1**		**1**		**1**		**1**	
3	**1**		**1**		**1**		**1**		**1**		**1**	
4	**1**		**1**		**1**		**1**		**1**		**1**	
5	**1**		**1**		**1**		**1**		**1**		**1**	
6	0	**1**	0	0	**1**		0		**1**		0	
7	0	**1**	0	0	**1**		0		**1**		0	
8	0	**1**	**1**	0	**1**		0		**1**		**1**	
9	0	**1**	0	0	0	**1**	0	0	**1**		0	
10	0	**1**	0	0	0	**1**	1	0	**1**		0	
11	0	**1**	0	**1**	0	**1**	0	0	**1**		0	
12	0	**1**	0	**1**	0	**1**	0	1	0	**1**	0	0
13	0	**1**	0	0	0	**1**	0	0	0	**1**	0	0
14	0	**1**	0	0	0	**1**	0	1	0	**1**	0	**1**
15	0	**1**	0	0	0	**1**	0	0	0	**1**	0	0
16	0	**1**	**1**	0	0	**1**	0	0	0	**1**	**1**	0
17	0	**1**	0	**1**	0	**1**	1	0	0	**1**	**1**	0
18	0	**1**	**1**	0	0	**1**	0	1	0	**1**	0	**1**
19	0	**1**	0	0	0	**1**	1	0	0	**1**	0	0
20	0	**1**	0	0	0	**1**	0	0	0	**1**	0	0
Σ	5.00	20.00	8.00	8.00	8.00	20.00	8.00	8.00	11.00	20.00	8.00	8.00
ρ	0.25		0.25		0.40		0.40		0.55		0.55	

Figure A3. Each entry again represents the mean of the distribution from which a feature-level memory-match signal is drawn (1 or 0, depending on whether the feature matches memory). These values are shown for three filler-similarity conditions (Low, Medium, and High). Manipulating filler similarity is conceptualized as manipulating the number of features not included in the witness's description (i.e., $f_6 \rightarrow f_{20}$) that match between faces in the lineup. The correlation between the overall (summed) memory signal of faces across lineups is determined the number of feature settings that match between faces in a lineup. For both TP and TA lineups in this hypothetical example, the number of features that match between two faces in the lineup is 5 in a low-similarity lineup (m = 5), 8 in a medium-similarity lineup (m = 8), and 11 in a high-similarity lineup (m = 11). The correlation (ρ) shown at the bottom is equal to m/n, where n = 20.

Modeling Multinomial Variability

The analyses presented in the main text assumed fixed settings for the binomial variables. For example, in medium similarity lineups, we have assumed that n = 20, n_D = 5, m = 8, and that $n_G = 20$ and $n_I = n_{F_{TA}} = n_{F_{TP}}$ = 8 for every lineup. The variability across lineups (yielding correlated summed memory signals) was entirely attributable to Gaussian error variance associated with the feature-level memory-match signals. Taking that approach makes it easy to mathematically work out the connection between feature-matching logic and signal detection logic. However, allowing for binomial variability would make the model more realistic.

To investigate the effect of allowing binomial variability, we simulated memory signals resulting from a world in which (1) faces are represented by 40 features instead of 20 and (2) the probability that a facial feature would be encoded is .50 (such that $\bar{n} = 20$). In this simulated world, if a facial feature of the perpetrator happened to be encoded, the probability that it would be included in the eyewitness's description was .25 (such that $\bar{n}_D = 5$), thereby appearing on every lineup member and also matching memory of the perpetrator. Additional facial features not included in the eyewitness's description were assumed to match memory of the perpetrator (and, independently, to match other faces in the lineup) with probability 1/5 = .2. Thus, the number of such features varied across lineups such that n_G, n_I, $n_{F_{TA}}$, and $n_{F_{TP}}$ (number of features that match memory) were all random variables, and so was m (the number of features that match across faces in the lineup).

Earlier, we noted in the absence of binomial variability, and before manipulating filler similarity, $\mu_G = n = 20$, $\mu_{\hat{G}} = n_D + .20(n - n_D) = 8$, and $\sigma = \sqrt{20} = 4.47$ (Figures 4 and 5). In that case, $m = 8$ and $\rho = m/n = .40$. When we allow for binomial variability, $\mu_G = \bar{n} = 20$, which is the same before, but the standard deviation of the distributions increases to $\sigma_G = 5.46$,

$\sigma_{\hat{G}} = 4.90$ (i.e., an unequal-variance model applies), and $\rho = .45$. Thus, the values change only slightly.

The same is true after manipulating filler similarity. To simulate high-similarity fillers, we could change the probability of a random match from .20 to .40 (more non-described features now match by design). When we do, $\mu_G = \bar{n} = 20$, as before, and the standard deviation of the distributions remain $\sigma_G = 5.46$, $\sigma_{\hat{G}} = 4.90$, However, now, $m = n_D + .40(n - n_D) = 11$ (as in Figure 4 for the high-similarity condition), and $\rho = .61$ (i.e., the correlation increases relative to the medium-similarity condition). To simulate low-similarity fillers, we could change the probability of a random match from .20 to 0 (fewer non-described features match than would match by chance). When we do, $\mu_G = \bar{n} = 20$, as before, and the standard deviation of the distributions remain $\sigma_G = 5.46$, $\sigma_{\hat{G}} = 4.90$, However, now, $m = n_D + 0(n - n_D) = 5$, and $\rho = .27$ (i.e., the correlation decreases relative to the medium-similarity condition). Thus, the specific values change somewhat when considerable binomial variability is introduced (e.g., the standard deviations and the correlations are all somewhat higher), but the basic patterns remain the same. Therefore, for the signal detection models we consider, we do not model the added complexity of allowing for binomial variability on the assumption that doing so would shed little additional light.

Model Flexibility

To investigate this issue, we generated simulated data from both models using the best-fitting parameter estimates shown in Table 1 and then fit both models to both sets of simulated data. We did this five times, with 3400 observations (i.e., simulated lineups) in each simulation, and the results are shown in Table A1. The data in the two left columns show fits of the two

models to data generated by the Ensemble model, and the data in the two right columns show fits of the two models to data generated by the Independent Observations model.

	Ensemble		Ind Obs	
Run #	Ens	Ind Obs	Ens	Ind Obs
1	9.4	52.1	180.8	20.0
2	32.1	47.9	153.5	11.4
3	20.0	30.1	115.3	24.1
4	18.8	68.6	165.2	11.1
5	21.9	43.8	209.4	14.7
mean χ^2	20.4	48.5	164.9	16.2
df	20	19	20	19

Table A1. χ^2 goodness-of-fit statistics based on maximum likelihood fits of the Ensemble model and Independent Observations model to five runs of simulated data generated by the Ensemble model (left to columns of data) and to simulated data generated by the Independent Observations model (right to columns of data).

Not surprisingly, the Ensemble model fits its own data well ($\chi^2 = 20.4$), but the Independent Observations model also does a reasonably good job of fitting the Ensemble model data ($\chi^2 = 48.5$). However, the reverse is not true. The Independent Observations model fits its own data well ($\chi^2 = 16.2$), but the Ensemble model does an abysmal job of fitting the Independent Observations model data ($\chi^2 = 164.9$). Thus, the fact that the two models fit the empirical data about equally well actually provides some support for the Ensemble model as well.

Computing Unequal-Variance Discriminability Measures

If σ were equal for filler and suspect distributions alike, then, for the Independent Observations model, the parameter estimates for μ_G and $\mu_{F_{TP}}$ can be used to calculate d'_{TP}, where $d'_{TP} = \frac{\mu_G - \mu_{F_{TP}}}{\sigma\sqrt{(1-\rho)}}$, as shown earlier in Equation 3. Setting σ to 1 and ρ to 0 (because that was its estimated value), the equation would reduce to $d'_{TP} = \left(\mu_G - \mu_{F_{TP}}\right)/\sigma$. Similarly, in that

simplified scenario, $d'_{TA} = (\mu_I - \mu_{F_{TA}})/\sigma$. However, given the specific design of the study (i.e., fixed suspects but variable fillers), σ_F turned out to be greater than 1. Thus, an unequal-variance discriminability measure would apply (often denoted d_a to distinguish it from d') in which the denominator would be the root mean square of σ and σ_F, or $\sqrt{.5(\sigma^2 + \sigma_F^2)}$, where σ is the standard deviation for innocent and guilty suspect distributions ($\sigma = 1$). We denote the relevant discriminability measures d_{TP} and d_{TA}. Thus, for the Independent Observations model, the expressions for d_{TP} and d_{TA} would be $d_{TP} = (\mu_G - \mu_{F_{TP}})/\sqrt{.5(\sigma^2 + \sigma_F^2)}$ and $d_{TA} = (\mu_I - \mu_{F_{TA}})/\sqrt{.5(\sigma^2 + \sigma_F^2)}$. The discriminability measure for innocent vs. guilty suspect (d'_{IG}) is a true d' score because the innocent and guilty suspect distributions have equal variance: $d'_{IG} = (\mu_G - \mu_I)/\sigma$, where $\sigma = 1$.

For the Ensemble model, the expressions for d_{TP} and d_{TA} would be $d_{TP} = \mu_{G-F_{TP}}/\sqrt{.5(\alpha\sigma^2 + \alpha\sigma_F^2)}$ and $d_{TA} = \mu_{I-F_{TA}}/\sqrt{.5(\alpha\sigma^2 + \alpha\sigma_F^2)}$, respectively, where $\alpha = 1 - 1/k$, and $k = 6$. For this model, d'_{IG} is also a true d' score because the innocent and guilty suspect distributions have equal variance: $d'_{IG} = (\mu_{G-F_{TP}} - \mu_{I-F_{TA}})/\sigma$, where $\sigma = 1$.