The Effects of Filler Similarity and Lineup Size on Eyewitness Identification

Kyros J. Shen, Jiaqi Huang, Allan L. Lam, & John T. Wixted

Department of Psychology, University of California, San Diego

Author Note

Correspondence concerning this article should be addressed to John T. Wixted (jwixted@ucsd.edu). The materials used in these experiments and the data reported here are available at https://osf.io/wv6tz/. Experiment 1 was not preregistered but Experiment 2 was (link: https://aspredicted.org/i4kj2.pdf).

Abstract

A photo lineup, which is a cross between an old/new and a forced-choice recognition memory test, consists of one suspect, whose face was either seen before or not, and several physically similar fillers. First, the participant/witness must decide whether the person who was previously seen is present (old/new) and then, if present, choose the previously seen target (forced choice). Competing signal-detection models of eyewitness identification performance make different predictions about how certain variables will affect a witness's ability to discriminate previously seen (guilty) suspects from new (innocent) suspects. One key variable is the similarity of the fillers to the suspect in the lineup, and another key variable is the size of the lineup (i.e., the number of fillers). Previous research investigating the role of filler similarity has supported one model, known as the Ensemble model, whereas previous research investigating the role of lineup size has supported a competing model, known as the Independent Observations model. We simultaneously manipulated these two variables (filler similarity and lineup size) and found a pattern that is not predicted by either model. When the fillers were highly similar to the suspect, increasing lineup size reduced discriminability, but when the fillers were dissimilar to the suspect, increasing lineup size enhanced discriminability. The results suggest that each additional filler adds noise to the decision-making process and that this noise factor is minimized by maximizing filler dissimilarity.

The Effects of Filler Similarity and Lineup Size on Eyewitness Identification

Photo lineups have largely replaced the live lineups once used by the police. A photo lineup consists of one suspect (who is either innocent or guilty) and several known-to-be-innocent fillers who are physically similar to the suspect. In years gone by, and sometimes still today, the outcome of a police lineup test was largely determined by various biasing factors instead of being determined by the witness's memory of the perpetrator (e.g., the use of an unfair lineup, or the lineup administrator steering the witness to the suspect). Since the late 1990s, science-based recommendations have been proposed to address issues like these (Wells et al. 1998, 2020). Examples include selecting fillers in such a way that the suspect does not stand out in the lineup, using at least five fillers, and having the lineup administered by someone who is blind to the identity of the suspect. When these and other recommendations are followed, a lineup procedure provides an objective test of eyewitness memory (National Research Council, 2014; Wells et al., 2020). Moreover, under such conditions, the results of standard laboratory-based research using fair lineups to elucidate the cognitive processes that underlie eyewitness identification decisions become relevant to real-world lineups.

In a lineup study conducted in the lab, participants first view the face of a "perpetrator" and are later presented with either a "target-present" (TP) or "target-absent" (TA) photo lineup. In a TP lineup, a photo of the perpetrator is surrounded by photos of the fillers. In a TA lineup, the perpetrator's photo is replaced by a photo of another filler who serves as an innocent suspect. The participant can either identify someone from the lineup (the suspect or a filler) or reject the lineup. The fillers play a key role in the effectiveness of lineup procedure, and the research we present here was designed to shed light on how and why they do.

Fillers in Photo Lineups

An eyewitness identification test without fillers is called a “showup.” In a showup, the singular suspect (innocent or guilty) is presented to the witness for a binary yes/no decision. Lineups have been consistently found to yield higher discriminability than showups (e.g., Akan et al., 2020; Neuschatz et al., 2016; Wetmore et al., 2015; Wooten et al., 2020). Empirically, higher discriminability means that witnesses are better able to sort innocent and guilty suspects into their respective categories. Theoretically, it means that the memory signal distributions generated by innocent and guilty suspects overlap to a lesser degree (Wixted & Mickes, 2018).

To have a beneficial effect on discriminability, the fillers cannot be randomly selected. For example, an unfair lineup with a black suspect and five white fillers would be unlikely to enhance discriminability compared to a showup because even someone who had not seen the perpetrator would be able to pick out the suspect. Therefore, fillers should be chosen in such a way that the suspect does not stand out (Wells et al., 2020). Doing so requires that the fillers have some degree of similarity to the suspect. But how similar should they be?

There are two common approaches to creating fair lineups: suspect matching and description matching. Using the suspect matching approach, the fillers are selected if their overall similarity to the suspect is subjectively judged to be sufficiently high. Using this approach, the police might select fillers who not only have the same race, age, and gender as the suspect but also have similar eyebrows, similar cheekbones, similar noses, etc. By contrast, using a description-matching approach, the fillers are chosen based on the description of the perpetrator provided by the witness, without comparing the fillers to the suspect at all (Luus & Wells, 1991). For example, if the witness described the perpetrator as a clean-shaven white male in his mid-30s, potential fillers who have those features would be suitable candidates for inclusion in the lineup.

A problem with the suspect-matching approach is that the task becomes impossible if the fillers are too similar to the suspect (Wells et al., 1993). Indeed, choosing fillers who are more similar than the degree of similarity achieved by the description-matching approach makes it harder to choose the guilty suspect in TP lineups without affecting the probability of choosing the innocent suspect in TA lineups (Colloff et al., 2021; Wells et al., 1993). Thus, it has been argued that the description-matching approach optimizes lineups by creating "propitious heterogeneity" (Wells et al., 1993).

Beyond description-matching, is there more that can be done to the filler-selection process to further enhance discriminability in fair lineups? Here, we focus on two manipulations that have been investigated independently and that have interesting theoretical implications: (1) reducing filler similarity and (2) increasing lineup size. With regard to reducing filler similarity, Colloff et al. (2021) found that selecting description-matched fillers who are otherwise *dissimilar* to the suspect increased discriminability compared to using medium- or high-similarity fillers. Consistent with an earlier report by Wells et al. (1993), the results showed that when description-matching is used (ensuring a fair lineup), dissimilar fillers yielded a higher hit rate than similar fillers while having no apparent effect on the false alarm rate. ROC analysis confirmed that discriminability increased as filler similarity decreased. In a follow-up study, Shen et al. (2023) directly manipulated filler similarity with face-morphing software and reported the same pattern of results. These findings suggest that by using a fair description-matched lineup, discriminability can be further enhanced by maximizing filler dissimilarity.

But how many fillers (dissimilar or otherwise) should be used? Although lineups have been shown to yield higher discriminability than showups (i.e., a lineup of size two is superior to a showup of size one), the number of fillers needed to optimize discriminability remains unclear. Two recent studies found that increasing lineup size beyond two did not further enhance discriminability (Akan et al., 2020; Wooten et al., 2020). This seems to suggest that a lineup size of two is sufficient, but there might be more to the story than that. Here, we manipulated lineup size using fillers who were either similar or dissimilar to the suspect in the lineup. We next consider predictions about what should be observed according to two signal detection models of lineup performance.

Theoretical Interpretation of the Role of Fillers in Photo Lineups

According to signal detection theory, a face in an eyewitness identification procedure (whether a showup or a lineup) generates a memory signal drawn from a distribution that is typically assumed to be Gaussian. As illustrated in Figure 1, in a TP showup (no fillers), the guilty suspect generates a memory signal randomly drawn from a distribution with mean μ_G and standard deviation σ, and in a TA showup, the innocent suspect generates a memory signal randomly drawn from a distribution with lower mean μ_I and standard deviation σ.

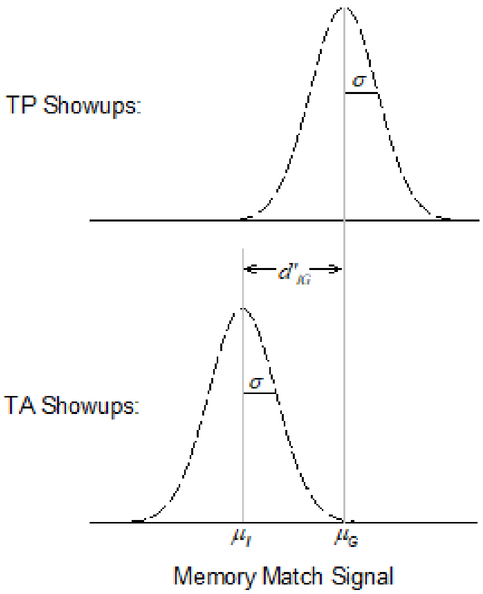

Figure 1. Target present (TP) and target absent (TA) memory match distributions for showups. An equal-variance model is assumed, so both distributions have the same standard deviation (σ). The means of the innocent and guilty suspect distributions are represented by μ_I and μ_G. The discriminability measure (d'_{IG}) is the standardized difference between the innocent suspect (TA) distribution and the guilty suspect (TP) distribution in showups.

Because a showup involves the presentation of a suspect without fillers, the only relevant discriminability measure in terms of underlying memory signals is d'_{IG}, which is the distance between the mean of the innocent suspect distribution and the mean of the guilty suspect distribution divided by their common standard deviation. That is, $d'_{IG} = \frac{\mu_G - \mu_I}{\sigma}$. As d'_{IG} increases, witnesses become better at distinguishing between innocent and guilty suspects. This measure is the same d' measure commonly used in studies of old/new recognition memory.

Unlike a showup, when memory is tested using a lineup, a witness has to contend with multiple memory signals before making a decision. Applying signal detection theory to lineups therefore requires additional assumptions, and competing models of lineup memory differ in what those assumptions are. The two main models considered to date are the Independent Observations model and the Ensemble model (Wixted et al., 2018, 2021), both of which can be used to interpret the effects of filler similarity and lineup size (i.e., the number of fillers) on discriminability.

Independent Observations model

According to the Independent Observations model, the memory signals generated by the suspect and the fillers in a lineup are considered at face value (i.e., in terms of their raw, untransformed memory signals drawn from Gaussian distributions), and they are independent of each other. For example, whether they are weak or strong, the memory-strength values drawn from the target distribution across lineups are not affected by (i.e., are independent of) the memory-strength values drawn from the filler distribution, and this is true no matter how many fillers there are in the lineup or how similar they might be to the target. In addition, the decision-making process is as simple as it could be: if the strongest (i.e., MAX) memory-match signal in the lineup (suspect or filler) exceeds a decision criterion, then that face is identified, and the stronger the MAX signal is, the more confident the eyewitness will be (independent of the strength of the memory signals generated by the other faces in the lineup). Note that the *decision variable* is the variable that is used to decide whether or not to make an ID and, if so, how confident the ID should be. Thus, if the Independent Observations model is correct, the raw (untransformed) memory signal associated with the MAX face is the decision variable.

As in a TP showup, in a TP lineup, the guilty suspect memory signal is conceptualized as a random draw from the guilty suspect distribution with the mean μ_G. In addition, each filler memory signal is conceptualized as a random draw from the filler distribution, which has a lower mean μ_{FTP}. The mean is lower because, unlike the guilty suspect, the fillers have not been seen before. Similarly, in a TA lineup, the innocent suspect memory signal is conceptualized as a random draw from the innocent suspect distribution with mean μ_I, and the memory signal for each filler memory signal is conceptualized as a random draw from the filler distribution with mean $\mu_{F_{TA}}$. In fair lineups, the innocent suspect distribution and the two filler distributions (with means of $\mu_{F_{TP}}$ and $\mu_{F_{TA}}$) are equivalent because the innocent suspect is just another filler, and the fillers are all matched to the basic physical characteristics of the perpetrator (so, on average, they should match memory to an equivalent degree). Thus, $\mu_I = \mu_{F_{TA}} = \mu_{F_{TP}}$ in the simplest case.

Figure 2 depicts memory-match distributions of suspects and fillers in target-present lineups and target-absent lineups. In addition to d'_{IG}, two new within-lineup discriminability measures become relevant when a lineup is used: d'_{TA} and d'_{TP}. When the memory signals generated by suspects and fillers in a lineup are uncorrelated, d'_{TA} is the standardized difference between the mean of the innocent suspect distribution and the mean of the filler distribution in TA lineups, and d'_{TP} is the standardized difference between the mean of the guilty suspect distribution and the mean of the filler distribution in TP lineups. In actual police lineups, however, the memory signals within a lineup are likely to be correlated because, by design, the faces in a lineup share features (e.g., they are all young white males who are clean-shaven with short dark hair). Under those conditions,

$$d'_{TA} = \frac{\mu_I - \mu_{FTA}}{\sigma\sqrt{1-\rho}} \tag{1}$$

and

$$d'_{TP} = \frac{\mu_G - \mu_{FTP}}{\sigma\sqrt{1-\rho}} \tag{2}$$

where ρ represents the degree to which the memory signals generated by the faces in the lineup are correlated with each other (Shen et al., 2023; Wixted et al., 2018, 2021). Note that we still refer to this as the Independent Observations model because, as we use the term, “independent” does not refer to statistical independence (ρ = 0). Instead, it refers to the fact that fillers in a lineup have no effect on memory strength for targets, and non-maximum memory strengths for a given lineup have no effect on the decision or confidence reported for that lineup.

Because the means of the innocent suspect and filler distributions are the same in a fair TA lineup, d'_{TA} is typically equal to 0. By contrast, d'_{TP} is greater than 0 because $\mu_G > \mu_{FTP}$.

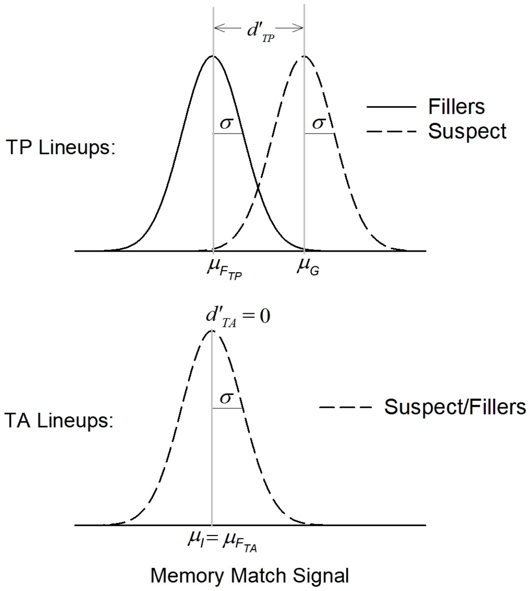

Figure 2. Memory match distributions for lineups. This is the same model as the one illustrated in Figure 1 except that a filler distribution has been added for TP lineups (whereas the filler distribution for TA lineups is the same as the innocent suspect distribution.

Ensemble model

Unlike the Independent Observations model, the Ensemble model holds that the operative memory signals are not the raw memory signals generated by the faces in the lineup. Instead, the raw signals are each transformed by subtracting away the mean memory signal associated with the faces in the lineup. This creates new distributions similar to those depicted in Figure 2 except that they represent distributions of difference scores. As with the Independent Observations model, the decision is still based on the MAX face. However, in this model, the lineup decision is based on the MAX minus mean difference score (i.e., the difference score is the decision variable). If the MAX-minus-mean decision variable exceeds a decision criterion, the MAX face will be identified, and the larger this MAX-minus-mean difference is, the more confident the eyewitness will be when making that ID. Thus, unlike the Independent Observations model, the decision about the MAX face is *not* independent of the strength of the memory signals associated with the other faces in the lineup.

In the Ensemble model, the equations for d'_{TA} and d'_{TP} are similar to those of the Independent Observations model except that they include lineup size (k):

$$d'_{TA} = \frac{\mu_I - \mu_{F_{TA}}}{\sigma\sqrt{(1-\rho)(1-1/k)}} \tag{3}$$

and

$$d'_{TP} = \frac{\mu_G - \mu_{F_{TP}}}{\sigma\sqrt{(1-\rho)(1-1/k)}} \tag{4}$$

The derivation of these equations can be found in prior work (Shen et al., 2023; Wixted et al., 2018, 2021).

Predictions regarding filler similarity

Both models agree that d'_{TA} should remain equal to 0 regardless of filler similarity because making description-matched fillers more or less similar to the innocent suspect should not make the innocent suspect provide a better or worse match to memory of the perpetrator, on average, compared to the fillers in the lineup (Colloff et al., 2021; Shen et al., 2023). Thus, it should always the be case that $\mu_I - \mu_{F_{TA}}$(the numerator of Equations 1 and 3) will equal 0.

Both models also agree that d'_{TP} should increase as filler similarity decreases (i.e., as $\mu_{F_{TP}}$ decreases). That is, as filler similarity decreases, $\mu_G - \mu_{F_{TP}}$ (the numerator of Equations 2 and 4) increases. This predicted effect is illustrated in Figure 3 using the Independent Observations model in the simplest (but unrealistic) case in which $\rho = 0$.

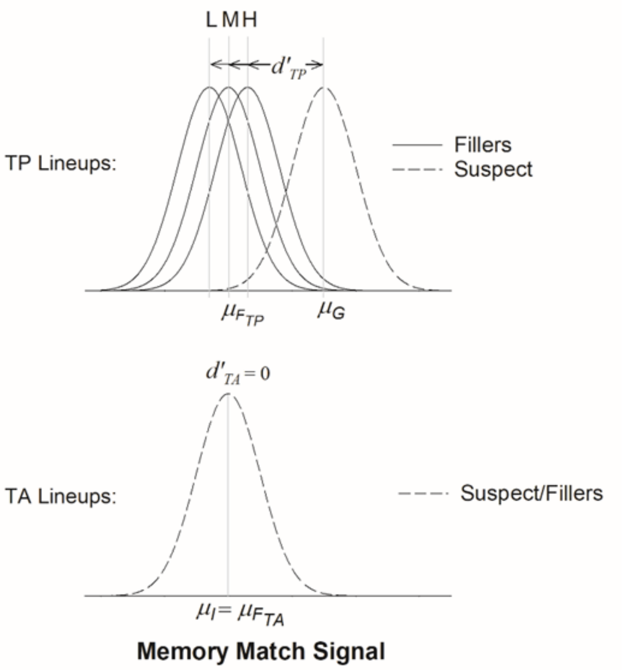

Figure 3. Distributions of memory signals in the low (L), medium (M), and high (H) filler-similarity conditions according to the Independent Observations model. These are the raw memory signals, before taking into account the effect of correlated signals.

The seemingly straightforward prediction illustrated in Figure 3 is complicated by the fact that as the fillers and the suspect become increasingly similar (minimizing the difference between μ_G and μ_{FTP}), the memory signals generated by the faces in the lineup become increasingly correlated, so ρ will not equal 0. In the extreme, when the faces become so similar that they are identical, ρ would equal 1.0. As shown in Equations 2 and 4, the increasing correlation with increasing filler similarity is a force that should increase d'_{TP}. Thus, the two opposing forces (i.e., $\mu_G - \mu_{FTP}$ decreasing but ρ increasing as filler similarity increases) make it difficult to intuitively infer what the models predict about manipulating filler similarity in TP lineups. Recently, however, Shen et al. (2023) showed that a simple feature-matching instantiation of both the Independent Observations model and the Ensemble model predicts that the negative force created by the mean signals becoming closer to each other outweighs the positive force of the increased correlation. Thus, according to that feature-matching version of the two models, it remains true that d'_{TP} should vary inversely with filler similarity.

Unlike the predicted effect of manipulating filler similarity on d'_{TA} and d'_{TP}, which is the same for the two models, the predicted effect on d'_{IG} differs for the two models. The Independent Observations model predicts that manipulating filler similarity should have no effect on d'_{IG}. That is, because memory signals in this model are independent of each other, manipulating the raw (i.e., untransformed) memory signals of the fillers should not affect the raw memory signals generated by the guilty suspect in TP lineups or the innocent suspect in TA lineups. Moreover, the concept of correlated memory signals does not apply to this discriminability measure because the innocent and guilty suspects appear in different lineups (whereas the correlation is a measure that applies to faces within lineups). Equation 5 presents the d'_{IG} equation for this model (Wixted et al., 2018):

$$d'_{IG} = \frac{\mu_G - \mu_I}{\sigma} \tag{5}$$

According to this equation (which is the same equation that applies to showups), the memory signals generated by the guilty suspect and the innocent suspect should be unaffected (and d'_{IG} should remain unchanged) when only the memory signals of fillers are manipulated by making them more or less similar to the suspect.

In contrast to the Independent Observations model, which operates on the raw memory match signals generated by the faces in the lineup, the Ensemble model operates on transformed memory signals, yielding different predictions regarding the effect of filler similarity. As described by Shen et al. (2023), for the Ensemble model, the equation for d'_{IG} is given by:

$$d'_{IG} = \frac{(\mu_G - \mu_{FTP})\sqrt{1 - 1/k}}{\sigma\sqrt{1 - \rho}} \qquad (6)$$

According to this equation, d'_{IG} (like d'_{TP}) should decrease with increasing filler similarity because μ_{FTP} will increase as filler similarity increases. The theoretical explanation is that the transformed guilty suspect distribution (based on a difference score) is directly affected by the memory strength of the fillers. Highly similar fillers generate strong absolute memory signals, which reduce the difference between the guilty suspect memory signal and the average memory signal strength of the lineup. Therefore, d'_{IG} decreases as filler similarity increases. The degree to which memory signals are correlated within a lineup (ρ) should also increase as μ_{FTP} increases, acting as a force to increase discriminability. As before, however, Shen et al. (2023) argued that this force is outweighed by the difference between the means of the suspect and filler distributions in TP lineups ($\mu_G - \mu_{FTP}$). Thus, despite the increasing correlation, d'_{IG} should decrease as filler similarity to the suspect increases.

The key point is that the two models make different predictions about how filler similarity should affect d'_{IG}. The Independent Observations model predicts that this discriminability measure should not vary as a function of filler similarity, whereas the Ensemble model predicts that it should. These predictions were evaluated in Shen et al. (2023) by fitting both models to empirical filler-similarity data, and the results of two experiments clearly supported the prediction of the Ensemble model. Indeed, even according to the fits of the Independent Observations model, d'_{IG}varied significantly as a function of filler similarity in the direction predicted by the Ensemble model (i.e., d'_{IG} varied inversely with filler similarity in description-matched lineups). The same conclusion emerged from a non-model-based approach of plotting detection ROCs, as we do here as well when presenting our results. A detection ROC treats any ID from a TP lineup (guilty suspect or filler) as a hit and any ID from a TA lineup (innocent suspect or filler) as a false alarm.

Lineup Size

The two models also make different predictions about how the manipulation of lineup size (k) should affect d'_{IG}. According to Equation 5, the Independent Observations model predicts that d'_{IG} should be independent of lineup size because k does not appear in that equation. Again, this is because the memory signals of the innocent and guilty suspects are independent of the memory signals associated with fillers, so it does not matter how similar the fillers are (as discussed above) or how many fillers there are. By contrast, according to Equation 6, the Ensemble model predicts that d'_{IG} should increase (with diminishing returns) as k increases.

As noted earlier, two recent studies found that empirical discriminability (measured by area under the ROC) increased when k increased from 1 (a showup) to 2 (a 2-person lineup) but did not increase further for lineups of $k = 3$ up to $k = 12$ (Akan et al., 2020; Wooten et al., 2020). The same was true when the two signal detection models were fit to the data to estimate d'_{IG} (Akan et al., 2020). Although both studies found a small trend in the direction predicted by the Ensemble model, there was no compelling evidence to reject the Independent Observations model. Studies of visual perception and ensemble coding have often generated conceptually analogous findings, where increasing set size led to a relatively constant sensitivity (Allik et al., 2013; Alvarez, 2011; Ariely, 2001; Chong & Treisman, 2005).

As an aside, it might seem as though the Independent Observations model cannot account for the increase in empirical discriminability (measured by area under the ROC) that is reliably observed as lineup size is increased from 1 (showup) to 2 (2-person lineup). However, it does allow for that effect because of the expected increase in d'_{TP} that occurs as a result of correlated memory signals in lineups (Equation 2), a consideration that does not apply to showups. However, beyond that initial effect of adding a filler to create a 2-person lineup, this model does not predict any further changes in discriminability as k increases, consistent with the empirical evidence.

This consideration brings up a potentially confusing point about the relationship between discriminability measures like d'_{IG} (the measure of primary interest in the work reported here) and an empirical measure of discriminability, such as partial area under the ROC curve (pAUC).[1] As noted by Wixted and Mickes (2018), although they often agree, the two measures of discriminability are dissociable, so if the question of interest concerns the effect of an experimental manipulation on d'_{IG} (e.g., as predicted by the Independent Observations model or the Ensemble model), one cannot always test that prediction by measuring the effect on pAUC. Instead, the effect on d'_{IG} must be assessed by fitting the relevant model to the data.

To illustrate this point, we simulated data for a showup and for lineups of size 2 and 6, first using the Independent Observations model and then using the Ensemble model. For these simulations, the standard deviation of all distributions was set to 1 ($\sigma = 1$), and the mean memory parameters were set to μ_G = 1.20 and $\mu_I = \mu_{FTA} = \mu_{FTP} = 0$. In other words, we generated the data based on a model like the one shown in Figure 2 for the Independent Observations model. For the Ensemble model simulation, we started with the same model and then simply transformed each simulated memory signal (suspect or filler) by subtracting away the mean memory signal for the lineup. In addition, three confidence criteria, c_1, c_2, and c_3 were

[1] The area under the ROC for lineups is a "partial" area because the false alarm rate range for a fair lineup (unlike a showup) is not 0 to 1 but is instead 0 to $1/k$, where k is lineup size.

set to 1.0, 1.5, and 2.0, respectively. For both models, we ran the simulation for lineup sizes of 1 (showup), 2, and 6, once with uncorrelated memory signals ($\rho = 0$) and once with correlated memory signals ($\rho = .60$). The simulated ROC data are shown in Figure 4. For each simulation for a given lineup size, there were 50,000 simulated TP trials and 50,000 simulated TA trials.

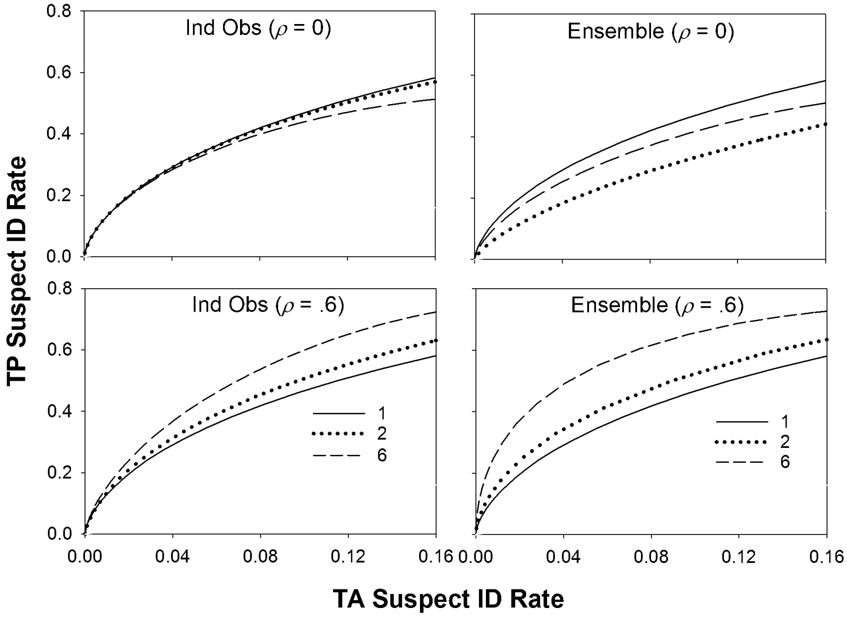

Figure 4. Simulated ROC data from the Independent Observations model (left graphs) and Ensemble model (right graphs). The graphs in the top row show simulated data with the correlation set to 0, whereas the graphs in the bottom row show simulated data with the correlation set to .60.

Consider first the simulated ROC data generated by the Independent Observations model (left two graphs in Figure 4). With the correlation set to 0 (top left graph), the showup ROC curve (lineup size = 1) and the two lineup ROC curves (lineup size = 2 and lineup size = 6) largely fall atop one another. This is an intuitively sensible result given that μ_G was the same for all three conditions (μ_G = 1.20), which means that d'_{IG} was the same for all three conditions as well. That is, according to Equation 5, $d'_{IG} = (1.20 - 0)/1 = 1.20$ for the showup and for both lineup size conditions. However, with the correlation set to .60 (bottom left graph), the showup ROC curve and the two lineup ROC curves now diverge, with the empirical area under the curve increasing as a function of lineup size. This is less intuitive given that μ_G was still the same for all three conditions (μ_G = 1.20), and d'_{IG} was the still same for all three conditions as well.

We next fit the simulated Independent Observations model data with the Independent Observations model using maximum likelihood estimation. Note that the model fit to the showup data is (here and elsewhere) a basic signal detection model, one that does not involve correlated memory signals or filler memory signals. That is, the model fit to the showup data was always the model shown in Figure 1. The natural expectation is that the models will fit the simulated data well and will also return the programmed parameter values. Table 1 shows the results, and those expectations are confirmed. That is, the parameter estimates correspond exactly to the programmed estimates (to the second decimal place), and the chi-square values indicate a very good fit. This exercise illustrates why it is essential to fit the Independent Observations model to the data to test its prediction that underlying discriminability remains constant as a function of lineup size. If the underlying memory signals are correlated, then relying an on pAUC to measure discriminability would misleadingly suggest that the Independent Observations model is wrong because that measure changes as a function of lineup size (Figure 4, lower left graph). By contrast, the parameter estimates obtained from fitting the model to data associated with correlated memory signals (and then computing d'_{IG}) yields the correct answer (Table 1). Note that the model fits also correctly recover the degree to which the memory signals are correlated.

Table 1. Maximum likelihood parameter estimates for μ_G, r, c_1, c_2, and c_3 when the Independent Observations model was fit to its own simulated data shown in Figure 4. The top three rows show estimates from the fit of the model to the uncorrelated simulated data, and the bottom three row show estimates from the fit to correlated simulated data. Also shown is d'_{IG} computed using Equation 5, the chi-square goodness-of-fit statistic, and the degrees of freedom (*df*), which is the degrees of freedom in the data minus the number of free parameters (4 parameters when *r* is fixed at 0 and 5 when it is free to vary). Note that the correlation parameter does not apply to a showup.

Lineup Size	d'_{IG}	μ_G	r	c_1	c_2	c_3	$\chi 2$	df
1	1.20	1.20	--	1.00	1.50	2.00	4.6	2
2	1.20	1.20	0.00	1.00	1.50	2.00	2.3	5
6	1.20	1.20	0.00	1.00	1.50	2.00	2.8	5
1	1.20	1.20	--	1.00	1.50	2.00	10.35	2
2	1.20	1.20	0.60	1.00	1.50	2.00	0.35	4
6	1.20	1.20	0.60	1.00	1.50	2.00	1.11	4

Now consider the simulated Ensemble model ROC data shown in Figure 4. In some ways, the results are surprising. For example, when the memory signals are uncorrelated (upper right graph), the lineup ROCs fall below the showup ROC. This means that using an Ensemble decision variable would be counterproductive if the lineup memory signals were uncorrelated. Even so, the model clearly predicts that as lineup size increases from 2 to 6, the ROC should increase. When the lineup memory signals are correlated (lower right graph), the simulated empirical ROC curve increases monotonically with lineup size. Indeed, whenever the correlation exceeds $1/k$ (where k is lineup size), the Ensemble model predicts that underlying discriminability will be enhanced relative to a showup.

We next fit the simulated Ensemble model data with the Ensemble model using maximum likelihood estimation (the showup data were again fit with the simple signal detection model shown in Figure 1). Once again, the models should fit the simulated data well and should also return the programmed parameter values, except there is one caveat for this model. As noted by Wixted et al. (2018), the likelihood function for the Ensemble model involves an approximation that is extremely accurate when the lineup size is 6, but the approximation is noticeably less accurate when the lineup size is only 2. Thus, we would expect the model to perform well only when the lineup size is 6 (or more). An additional consideration is that, when the data are based on correlated memory signals, the Ensemble model cannot recover the programmed correlation parameter. As explained by Wixted et al. (2018), the MAX – mean decision variable envisioned by this model subtracts away shared variance, leaving no information about the correlation contained in the confidence rating data. Yet, according to Equation 5, the magnitude of the correlation is a determinant of d'_{IG} for the Ensemble model. The implication is that when fitting this model to empirical data, the estimated d'_{IG} values are monotonically correct, but one would need to know the underlying correlation for the values to be exact. For the simulated data, the underlying correlation is known, and we used that information to compute d'_{IG} values based on the fits of this model.

Table 2 shows the results of the Ensemble model fits. When the lineup size is 6, the parameter estimates correspond almost exactly to the programmed estimates (to the second decimal place), and the chi-square values indicate a very good fit. However, as expected, when the lineup size is 2, the parameter estimates are imperfect, and the fit is no longer extremely good. Note that, unlike the fits of the Independent Observations model, the estimates of d'_{IG} based on the fits of the Ensemble model largely correspond to the empirical ROC data shown in Figure 4. That is, as a general rule, for the Ensemble model, d'_{IG} and pAUC tend to go hand in hand (see also Shen et al., 2023).

Table 2. Maximum likelihood parameter estimates for μ_G, c_1, c_2, and c_3 when the Ensemble model was fit to its own simulated data shown in Figure 4. Note the μ_G estimate is actually an estimate of $\mu_G - \mu_{FTP}$, but μ_{FTP} was set to 0 in these simulations. The top three rows show estimates from the fit of the model to the uncorrelated simulated data, and the bottom three row show estimates from the fit to correlated simulated data. Also shown is d'_{IG} computed using Equation 6, the chi-square goodness-of-fit statistic, and the degrees of freedom (*df*). Note that the correlation parameter does not apply to the showup, and although it is relevant to the two lineups, it cannot be estimated when the Ensemble model is fit to the data.

Lineup Size	d'_{IG}	μ_G	c_1	c_2	c_3	$\chi 2$	*df*
1	1.20	1.20	1.00	1.50	2.00	4.7	2
2	0.81	1.14	0.97	1.45	1.95	49.5	5
6	1.09	1.20	1.01	1.50	1.99	7.0	5
1	1.20	1.20	1.00	1.50	2.00	2.3	2
2	1.28	1.14	0.96	1.47	1.96	17.8	5
6	1.73	1.20	1.00	1.50	2.00	2.0	5

The key point of these simulations is that with the untransformed underlying memory distributions held constant, increasing lineup size should have no effect on d'_{IG} according to the Independent Observations model, but d'_{IG} should increase as a function of lineup size according to the Ensemble model. How d'_{IG} should compare to a showup according to the Ensemble model depends on the degree to which memory signals are correlated in a lineup. But whether they are correlated or not, d'_{IG} should increase as lineup size increases beyond $k = 2$.

Because of the complications associated with fitting the Ensemble model to lineups of size 2, we focus mainly on fits of the Independent Observations model. The outcome of these fits can test the predictions made by both models. For example, Shen et al. (2023) reported simulations involving manipulations of filler similarity. As with lineup size, the Independent Observations model predicts that d'_{IG} should remain constant as a function of filler similarity (i.e., d'_{IG} should not be affected by either the similarity or the number of fillers in the lineup). By contrast, the Ensemble model predicts that d'_{IG} should increase as filler similarity decreases. When the Independent Observations model was fit to simulated data generated by the Ensemble model, its estimates of d'_{IG} did not remain constant but instead varied as a function of filler similarity in the manner predicted by the Ensemble model (increasing as filler similarity decreased). The same d'_{IG} pattern was observed when the Independent Observations model was fit to empirical filler similarity data. Thus, the filler-similarity model-fitting results supported a priori predictions made by the Ensemble model.

Here, we show that similar considerations apply to predictions about the effect of lineup size on d'_{IG}. Specifically, we fit the Independent Observations model to the simulated data generated by the Ensemble model shown in lower right panel of Figure 4 (correlated memory signals). The best-fitting model yielded estimates of d'_{IG} (and, equivalently, μ_G) of 1.20 for the showup, 1.43 for the 2-person lineup, and 1.60 for the 6-person lineup. Thus, once again, the Independent Observations model can be fit to either filler-similarity data or to lineup-size data to test predictions about d'_{IG} made by both models.

Returning to the main point, research on the effect of filler similarity has favored the Ensemble model, whereas research on the effect of lineup size has favored the Independent Observations model. What explains the apparent discrepancy? Here, we address that question by simultaneously manipulating both filler similarity and lineup size. We did not have an a priori reason to believe that these two variables would interact. However, given the inconsistent verdict from those two separate lines of research, it seemed worthwhile to investigate the possibility that they do. In Experiment 1, we investigated the role of lineup size for high-similarity fillers, and in Experiment 2, we investigated the role of lineup size for low-similarity fillers.

Experiment 1

Beginning with fillers who were matched to the suspect on basic physical characteristics (in terms of race, gender, and age), in Experiment 1, morphing software was used to further increase similarity between the suspect and the fillers. Two levels of higher-than-normal similarity were achieved by morphing the suspect's face onto the fillers to two different degrees (namely, 20% vs. 60%). In addition, three lineup sizes were used (1, 2, and 6).

Method

Participants

In total, 1712 participants (M_{age}=34.12) were recruited through Amazon Mechanical Turk. Participants were included only if they successfully answered both 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 52.4% male (897), 46.7% female (800), 0.4% other (7) and 0.4% prefer not to state (8), with the ethnicity distribution being: 7.5% African-American (128), 16.8% Asian (287), 2.5% Mexican-American (43), 0.9% Filipino (16), 6.8% Latino (117), 1.9% Native-American (33), 57.7% Caucasian (988), 4.2% Other/Undeclared (72), 1.6% Prefer not to state (28). The experiments reported here were approved by the UC San Diego Institutional Review Board.

Design and Materials

We used a 2 (filler similarity: similar vs. highly similar) × 2 (lineup type: target-present vs. target-absent) × 3 (lineup size: showup/2-person lineup/6-person lineup) mixed factorial design. Filler similarity was a between-subject factor, while lineup type and lineup size were within subject factors. We selected all photos categorized as "white, male" from the Chicago Face Database (CFD; Ma, Correll, & Wittenbrink, 2015). The photos were randomly divided into six sets, each consisting of 15 faces, one of which was randomly selected and designated to serve as the suspect while the other 14 served as potential fillers.

To manipulate filler similarity, we altered the 14 potential fillers in each set with Fantamorph software to create two pools of photos to reflect the two similarity conditions: similar and highly similar. For the “similar” pool, the 14 fillers were morphed with the suspect to create new faces that were 20% suspect and 80% filler. For the “highly similar” pool, the same fillers were used to create new faces that were 60% suspect and 40% filler. As a result, we obtained six photo sets, each of which contained one designated suspect and two pools of fillers that resembled the suspect at two different levels, which we refer to as similar (20% suspect) and highly similar (60% suspect). Examples are shown in Figure 5.

The six sets of photos were randomly assigned to create the six lineups to reflect the six possible combinations of within-subject conditions: one target-present and one target-absent lineup for each of the three lineup sizes, showup, 2-person lineup and 6-person lineup. Each lineup included the designated suspect and a certain number (0, 1, 5) of fillers picked from the pool of 14, depending on the assigned lineup size. Since the lineups always included the designated suspect, whether they were TP or TA was dependent on the photo shown during the study phase. 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. This design is identical to the paradigm from Shen et al. (2023) except for the varied number of fillers. The face stimuli were also cropped into oval shapes. Doing so reduces real-world generalizability but has the advantage of making it easier to create morphed faces that do not look morphed. Moreover, the main goal of our study is to test theory-based predictions, and cropping the faces theoretically should not affect our conclusions. Manipulating methodological details that should not matter theoretically is a useful way to test the robustness of a model (e.g., Baribault et al., 2018).

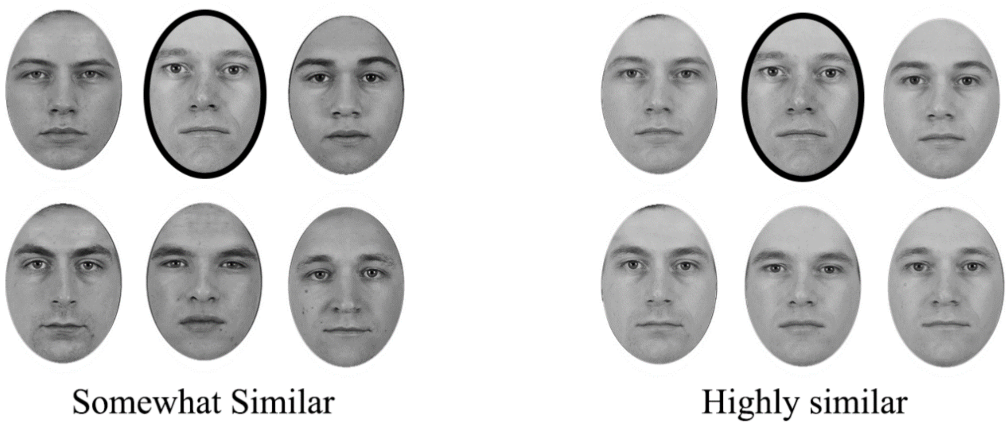

Figure 5. Examples of lineups constructed with stimuli at two similarity levels, "similar" (morphed with 20% suspect) and "highly similar" (morphed with 60% suspect).

Procedure

Each participant received six study-test trials. Every trial included a study phase, a 60-seconds distractor task, and a test phase. During each study phase, the participant viewed one photo for 3 seconds. After viewing the photo, the participant was given a distractor task: playing one of the two mini games, "Tetris" or "2048", for 60 seconds. The participant then viewed a single photo, a two-photo lineup or two-row by three-column photo lineup depending on the assigned condition, each with a 'Not Present' option underneath it. The spatial location of each photograph was randomized. The participant was given the instruction "Please choose the face you saw. If you do not recognize any of the faces, click on the 'Not Present' option." 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.

Experiment 1 was not preregistered. The materials used in this experiment and the data reported next are available at https://osf.io/wv6tz/ (Shen & Wixted, 2023).

Results and Discussion

The proportions and frequency counts of response outcomes (suspect ID, filler ID, no ID) for TP and TA lineups across three different lineup sizes and two similarity levels are shown in Table 3, and the corresponding ROC curves are shown in Figure 6. Note that the showups (lineup size=1) in the two similarity conditions are identical since showups include no fillers. Even so, their data are presented separately because they were collected from the participants assigned to the two different similarity conditions.

Table 3. Frequency counts (top) and proportions (bottom) of Suspect IDs, Filler IDs, and No IDs in the showup (lineup size = 1), 2-person and 6-person lineup conditions for Target-Present (TP) and Target-Absent (TA) lineups with similar and highly similar fillers. The "--" symbols represent nonexistent filler data for showups.

		TP counts			TA counts		
Similarity	Size	Suspect ID	Filler ID	No ID	Suspect ID	Filler ID	No ID
Similar	1	459	--	399	119	--	739
	2	436	21	401	107	72	679
	6	362	98	398	72	209	577
Highly similar	1	389	--	465	78	--	776
	2	347	98	409	93	102	659
	6	229	227	398	61	268	525

		TP proportions			TA proportions		
Similarity	Size	Suspect ID	Filler ID	No ID	Suspect ID	Filler ID	No ID
Similar	1	0.53	--	0.47	0.14	--	0.86
	2	0.51	0.02	0.47	0.12	0.08	0.79
	6	0.42	0.11	0.46	0.08	0.24	0.67
Highly similar	1	0.45	--	0.54	0.09	--	0.9
	2	0.4	0.11	0.48	0.11	0.12	0.77
	6	0.27	0.26	0.46	0.07	0.31	0.61

The ROC curves shown in Figure 5 reveal an apparent interaction between the effect of lineup size and filler similarity. In the Similar condition, the ROC curves for the three lineup sizes essentially fall on top of each other. In other words, there is no appreciable difference between the discriminability of the showups, 2-person lineups and 6-person lineups. For the 2- and 6-person conditions, this result is to be expected because our similar condition (20% morph to the suspect) differs only slightly from a typical lineup procedure in which all of the fillers match the basic description of the perpetrator without otherwise being matched in terms of similarity. Under those typical conditions, increasing lineup size beyond 2 has been found to have no measurable effect on discriminability (Akan et al., 2020; Wooten et al., 2020). However, prior research has found that a lineup size of 2 yields higher discriminability than a showup. That effect was not observed here in the Similar condition.

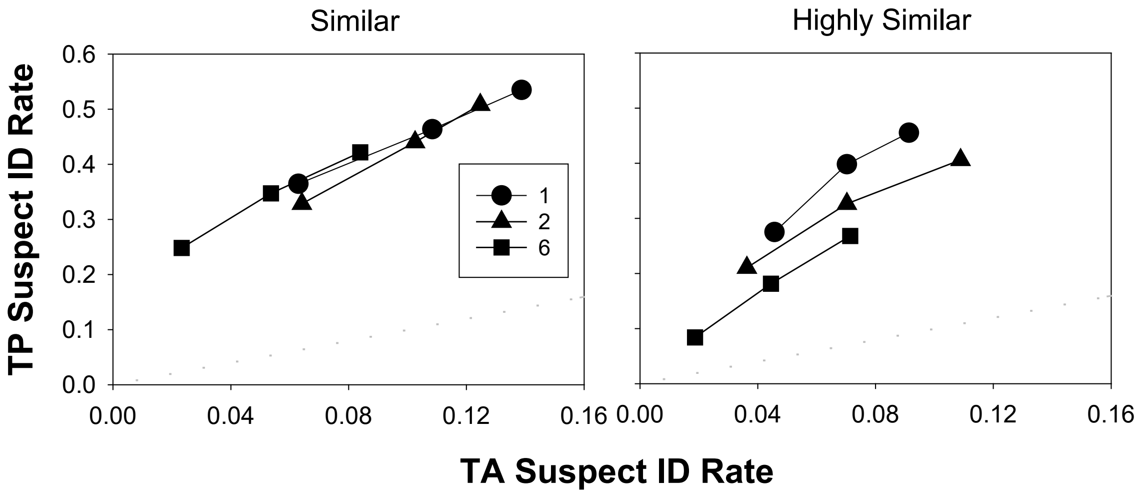

Figure 6. ROC data from the similar and highly similar conditions of Experiment 1. The TP suspect ID rate is the proportion of TP lineups that resulted in a suspect ID. The TA suspect ID rate is the proportion of TA lineups that resulted in a filler ID divided by the lineup size of 6 (a standard approach when fair lineups are used). Note that filler IDs from TP lineups are not represented in these ROC plots, but they are taken into account when models are fit to the data. The dashed line represents chance performance.

In the Highly Similar condition, showups actually yielded the highest level of discriminability, whereas 6-person lineups yielded the lowest level of discriminability. In other words, in this condition, the data showed an unexpected *negative* effect of lineup size on discriminability. To test the significance of these results, we fit the Independent Observations model to the data using maximum likelihood estimation.[2] Initially, the model was fit to the data from all six conditions simultaneously, with the free parameters either fixed or free to vary across conditions in a manner consistent with what the model naturally predicts. We then relaxed those assumptions, allowing certain parameters to vary that, according to this model, should not be affected by our experimental manipulations. Once that step is taken, it is technically no longer the Independent Observations model, and at that point, we are using signal detection theory as a measurement tool. This is the most basic and generic signal detection model for lineups, the one that would be used to quantify discriminability had neither the Independent Observations model nor the Ensemble model been proposed. It is mathematically equivalent to the Independent Observations model but is agnostic about whether the memory signals for suspects and fillers are independent.

For the initial fits, μ_I was set to 0 and μ_G was free to vary but was constrained to be equal across both filler similarity conditions and all three lineup size conditions (one parameter). These constraints reflect the fact that, according to the Independent Observations model, the memory signals generated by innocent and guilty suspects should be unaffected by the memory signals associated with the fillers (regardless of how many or how similar they are to the suspect). We set σ to 1 (i.e., an equal-variance model was assumed), and allowing it to vary never significantly improved the fit.

In addition, μ_{FTP} and μ_{FTA} were allowed to vary as a function of filler similarity because filler similarity was experimentally manipulated in both TP and TA lineups. The manipulation

[2] The model-fitting analyses assume independence, which is not strictly true because each participant contributed 6 observations instead of only 1.

would be expected to affect μ_{FTP} (it should be higher in the highly similar condition) but not necessarily μ_{FTA}. Even so, μ_{FTA} was free to vary across the two filler similarity conditions because filler similarity was experimentally manipulated in TA lineups, and that manipulation might have an effect (e.g., due to an unforeseen nuisance factor associated with morphing the fillers with the suspect). These parameters were not free to vary as a function of lineup size (2 vs. 6) because the mean of the filler memory signals should not change as a function of how many fillers are used. Because μ_{FTP} and μ_{FTA} were allowed to differ from each other within both filler-similarity conditions and to also differ across filler-similarity conditions, it added four additional free parameters.

Next, the correlation parameter (ρ) was free to vary as a function of filler similarity, adding two more parameters. The correlation between lineup memory signals should be higher in the Highly similar condition compared to the Similar condition. Finally, the three confidence criteria were free to vary, separately for each of the six conditions (18 parameters).[3] Altogether, the model involved 1 + 4 + 2 + 18 = 25 free parameters.

For model-fitting purposes, the ratings for lineup and showup rejections were aggregated across confidence levels. The decision variable upon which confidence for positive identifications is presumably the memory signal associated with the MAX face (either its raw memory signal or its transformed memory signal). However, when a lineup is rejected, no particular face is identified (i.e., the set of faces is rejected), and it remains an open question as to whether, in that case, confidence is still based on the MAX memory signal or is instead based on a collective memory signal (Lindsay et al., 2013; Weber and Brewer, 2006; Yilmaz et al., 2022).

[3] The confidence scale was collapsed into three levels (0-60, 70-80, and 90-100), thereby creating bins with similar numbers of observations and reducing the number of free parameters.

Because there are 60 degrees of freedom in the data and 25 free parameters in the model fit, there were 60 – 23 = 37 remaining degrees of freedom. Thus, the expected chi square goodness-of-fit given the right model is 37. The actual fit was $\chi^2(37) = 48.5, p = .01$, which means that the data did not quite deviate significantly from the predictions of the model. Table 4 shows the relevant d' values computed from the parameter estimates (also shown in the table) using Equations 1 through 3 presented earlier.

Table 4. Discriminability measures ($d'_{IG}, d'_{TP}, d'_{TA}$) and parameter estimates ($\mu_G, \mu_{FTP}, \mu_{FTA}, c_1, c_2, c_3$) for each lineup size ($k$) condition based on a fit of the Independent Observations model to the data from Experiment 1.

Similarity Condition	k	d'_{IG}	d'_{TP}	d'_{TA}	μ_G	μ_{FTP}	μ_{FTA}	c_1	c_2	c_3
Somewhat Similar	1	1.07	--	--	1.07	--	--	1.02	1.19	1.45
	2	1.07	1.67	0.33	1.07	-0.61	-0.33	1.07	1.23	1.51
	6	1.07	1.67	0.33	1.07	-0.61	-0.33	1.23	1.48	1.80
Highly Similar	1	1.07	--	--	1.07	--	--	1.24	1.38	1.69
	2	1.07	1.01	0.00	1.07	0.06	0.00	1.17	1.43	1.82
	6	1.07	1.01	0.00	1.07	0.06	0.00	1.47	1.80	2.23

Note. Estimates of d'_{TP}, d'_{TA}, μ_{FTP}, and μ_{FTA} apply only to 2- and 6-person lineup conditions.

In some ways, the estimates are reasonable. For example, d'_{TP} should decrease as filler similarity increases, and as expected, its estimated value is lower in the Highly similar condition (1.01) compared to the Similar condition (1.67). There was also a small effect on d'_{TA} that was not predicted but is perhaps not surprising given that filler similarity was experimentally manipulated in TA lineups as well. Curiously, however, the correlation estimate was close to 0 in the Similar condition and in the Highly Similar condition, and constraining ρ to equal 0 in both filler-similarity conditions did not significantly worsen the fit. This seems curious because the memory signals in lineups should be correlated, and the correlation should increase as filler similarity increases. Similar curiosities were reported by Shen et al. (2023).

We next tested a key prediction by allowing μ_G to vary for showups vs. lineups. That is, instead of using only one estimate of μ_G for all six conditions, we used one estimate for the two showup conditions (which were methodologically identical to each other) and another estimate for both the 2-person and 6-person lineup sizes. The addition of this one additional parameter resulted in a significant improvement of the fit, $\chi^2(1) = 4.55, p = .03$. As shown in Table 5, the estimates of d'_{IG} for lineups was reduced compared to the showup condition.

Table 5. Discriminability measures and parameter estimates based on a fit of the Independent Observations model to the data from Experiment 1, now with μ_G free to differ between the showup conditions and the lineup conditions.

Similarity Condition	k	d'_{IG}	d'_{TP}	d'_{TA}	μ_G	μ_{FTP}	μ_{FTA}	c_1	c_2	c_3
Somewhat Similar	1	1.17	--	--	1.17	--	--	1.08	1.25	1.52
	2	1.01	1.65	0.37	1.01	-0.64	-0.37	1.04	1.20	1.48
	6	1.01	1.65	0.37	1.01	-0.64	-0.37	1.20	1.44	1.76
Highly Similar	1	1.17	--	--	1.17	--	--	1.30	1.45	1.76
	2	1.01	1.00	0.00	1.01	0.01	0.00	1.15	1.41	1.79
	6	1.01	1.00	0.00	1.01	0.01	0.00	1.45	1.78	2.21

Finally, we allowed μ_G for the lineups to differ between the Similar and Highly similar conditions. The addition of this parameter improved the fit considerably, $\chi^2(1) = 13.4, p < .001$. As shown in Table 6, d'_{IG} is differentially reduced relative to showups in the Highly similar condition (model fits of the ROC data are presented in the Appendix).

Table 6. Discriminability and correlation estimates based on a fit of the Independent Observations model to the data from Experiment 1, now with d'_{IG} free to vary as a function of both filler similarity and lineup size.

Similarity Condition	k	d'_{IG}	d'_{TP}	d'_{TA}	μ_G	μ_{FTP}	μ_{FTA}	c_1	c_2	c_3
Somewhat Similar	1	1.18	--	--	1.18	--	--	1.09	1.26	1.53
	2	1.15	1.70	0.28	1.15	-0.56	-0.28	1.12	1.28	1.57
	6	1.15	1.70	0.28	1.15	-0.56	-0.28	1.28	1.53	1.85
Highly Similar	1	1.18	--	--	1.18	--	--	1.30	1.45	1.76
	2	0.93	0.92	0.00	0.93	0.01	0.00	1.12	1.37	1.75
	6	0.93	0.92	0.00	0.93	0.01	0.00	1.43	1.76	2.19

The ROC data in Figure 6 suggest that empirical discriminability for the 6-person lineups may be lower than that of the 2-person lineups in the Highly Similar condition. Allowing μ_G to vary as a function of lineup size in that condition reflected that trend ($\mu_G = 0.96$ for 2-person lineups vs. $\mu_G = 0.91$ for 6-person lineups), but it did not significantly improve the fit. Still, we interpret these results to mean that increasing lineup size impairs discriminability when highly similar fillers are used.

These results are inconsistent with the predictions of the Independent Observations model. That is, in all six conditions, the innocent and guilty suspects were the same, and according to this model the underlying memory signals they generate are *independent* of the memory signals generated by the other faces in the lineup. Therefore, manipulating the number and/or similarity of the fillers should not affect the means of the innocent or guilty suspect distributions. Yet d'_{IG} varied as a function of filler similarity in the manner predicted by the Ensemble model (decreasing as filler similarity increased) and also varied as a function of lineup size, with d'_{IG} being significantly *lower* for the two lineup conditions compared to the showup condition.

If lineup memory signals are uncorrelated, the Ensemble model can predict that the two lineup conditions would yield lower d'_{IG} scores compared to the showup condition (based on the simulations reported earlier). However, even in the uncorrelated case, it unambiguously predicts that d'_{IG} should be larger for the 6-person lineup than the 2-person lineup. Instead, the estimates did not differ, and the trend was in the opposite direction (i.e., lower d'_{IG} for the 6-person lineup). Moreover, it seems unlikely that the correlation was close to 0 given that the fillers are highly similar to the suspect. Instead, given how similar the faces were in the highly similar condition, the memory signals were presumably highly correlated. Thus, these results are not consistent with the Ensemble model either.

We did not fit the Ensemble model to the data in part because, as noted earlier, its estimates for the 2-person lineup would be imprecise. In addition, the fact that the lineup size trend was in the opposite direction to what this model predicts already establishes that it cannot accommodate these results. However, to illustrate how the results favor the Ensemble model over the Independent Observations model with respect to filler similarity, it is useful to also examine the detection ROCs. 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 Independent Observations and Ensemble models make qualitatively different predictions about the order of the ROCs across the two filler-similarity conditions.

According to the Independent Observations model, in a TP lineup, making the fillers more similar to the suspect should increase the chances that a face exceeds the decision criterion (elevating the hit rate). The reason is that on trials in which the target happened to be the MAX face but did not exceed the decision criterion, a similar filler independently drawn from the TP filler distribution with a high mean might now be the MAX face and also exceed the decision criterion. No such effect would be expected in TA lineups because making the fillers more similar to the innocent suspect should not affect their average memory strength. Thus, the detection ROC in the Highly Similar condition should exceed that of the Similar condition.

The Ensemble model makes the opposite prediction. Although it remains true that the manipulation of filler similarity should have no effect on the false alarm rate for TA lineups, the hit rate should *decrease* (not increase) as filler similarity increases. The reason is that the difference between the MAX face and the average of the lineup faces will be smaller when filler similarity is high, and it is that difference score that needs to exceed the decision criterion for a positive ID to be made. Thus, the detection ROC in the Similar condition should exceed that of the Highly Similar condition.

Figure 7 shows the detection ROCs for Experiment 1, and it is clear that the filler similarity effect predicted by the Independent Observations model is not observed. Instead, as predicted by the Ensemble model, for both the 2-person and 6-person lineups, the detection ROC for the Similar condition exceeds that of the Highly Similar condition. The same trends were reported by Shen et al. (2023). For these detection ROCs, both models predict that the ROC will decrease as lineup size increases, but neither model predicts that effect for the detection-plus-identifications ROCs for the Highly Similar condition shown in Figure 6.

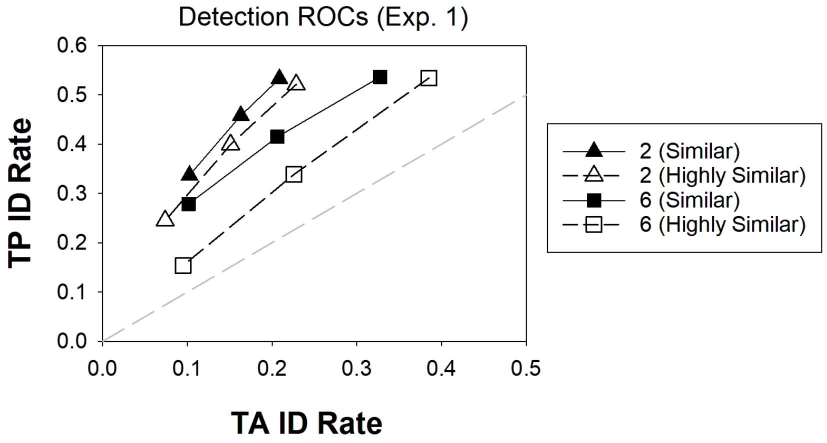

Figure 7. Detection ROC data from the similar and highly similar lineup conditions of Experiment 1. The TP ID rate is the proportion of TP lineups that resulted in a positive ID of a suspect or filler, and the TA ID rate is the proportion of TA lineups that resulted in a positive ID of a suspect or filler. The dashed line represents chance performance.

Overall, the results of Experiment 1 were somewhat unexpected and were not fully predicted by either model. For some reason, an effect of lineup size on d'_{IG} was only present in the highly similar condition, and d'_{IG} was not ordered in a way that could be predicted by either model. In other words, for reasons unknown, filler similarity appears to play a role in determining the effect of lineup size on discriminability.

Experiment 2

The two lineup conditions of Experiment 1 involved fillers that were more similar to the suspect than would be the case in a standard procedure involving description-matched fillers. This raises an interesting question: How would lineup size affect discriminability if the fillers were *less* similar to the suspect than would be the case in a standard procedure involving description-matched fillers? The two lineup conditions in Experiment 1 were created by morphing the face of the suspect to the fillers in the lineup, thereby ensuring that the fillers in both conditions had above-average similarity. When manipulating filler similarity in that direction, as filler similarity increases, the faces become ever more similar to a singular face (the suspect's face). Thus, in order to test the lineup size effect when fillers have below-average similarity, a different approach was needed. No longer can we morph the suspect's face onto the fillers to manipulate filler similarity because that approach always increases similarity relative to the starting point.

To manipulate filler similarity in Experiment 2, we created two new pools of fillers. This time, instead of morphing the fillers directly with the suspect, we morphed the fillers (at 60%) with faces that were independently rated as being similar to the suspect (the similar condition) vs. highly dissimilar to the suspect (the dissimilar condition). We could have directly used the rated fillers to create the two filler-similarity conditions, but we retained the face morphing approach to maintain compatibility with Experiment 1.

The similar condition in Experiment 2, like the similar condition in Experiment 1, involved fillers who were somewhat more similar to the suspect than would be the case had unmodified description-matched fillers been used (because the description-matched fillers were morphed with other fillers independently rated to be similar to the suspect). Although we used the same name in both experiments (the similar condition), the somewhat higher-than-normal filler similarity was achieved in different ways. Still, the conditions were comparable, so we expected comparable results as well.

The fillers in the dissimilar condition were morphed to other faces that had been rated as being dissimilar to the suspect. However, they were not so dissimilar as to create an unfair lineup. A basic requirement of a fair lineup–one that we followed here–is that everyone be matched on basic physical characteristics like age, race, and gender associated with the target in memory. In actual police investigation, this is best accomplished by ensuring that everyone in the lineup match the physical description of the perpetrator provided by the eyewitness (Wells et al., 1993). Matching lineup members on basic physical characteristics ensures that the lineup will always be fair (i.e., the suspect will not stand out) and that there will always be a non-trivial degree of similarity between the suspect (innocent or guilty) and the fillers in a given lineup. For purposes of Experiment 2, the important point is that the fillers in the dissimilar condition are considerably less similar to the suspect than would otherwise be the case, but they still match the basic characteristics of the suspect on the dimensions of age, race, and gender. Experiment 2 involved two filler similarity conditions (similar and dissimilar) and two lineup sizes (2 vs. 6).

Participants

Experiment 2 involved an initial rating study to identify fillers who had low or high similarity relative to suspect photos, followed by the main experiment. In total, 642 participants were recruited from the UC San Diego SONA system for the initial rating study. All participants were students attending UC San Diego at the time. For the main experiment, 1748 participants (M_{age}=34.8) 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 49.1% male (858), 49.7% female (869), 0.1% other (17) and 0.1% prefer not to state (4), with the ethnicity distribution being: 9.3% African-American (163), 9.6% Asian (167), 3% Mexican-American (53), 0.6% Filipino (11), 4% Latino (68), 7% Native-American (122), 62.4% Caucasian (1090), 2.9% Other/Undeclared (50), 1.4% Prefer not to state (24).

Design and Materials

A pool of 376 faces was rated (92 white male, 82 black male, 89 white female, 103 black female) and then used in the main experiment. In the initial rating study, each participant rated 52 faces (13 white male, 13 black male, 13 white female, 13 black female) randomly chosen from the 376 faces in terms of how much they resemble the single suspect face. Each individual face received 80 to 100 ratings, and an average score was calculated for each face. The five faces with the lowest average similarity score were morphed with fillers at 60% to create fillers for the dissimilar condition, and the five faces with the highest average similarity score were morphed with other fillers at 60% to create fillers for the similar condition.

Note that this manipulation is different from that of Experiment 1 because, in Experiment 2, each filler in the lineup was morphed with a *different* face (e.g., 5 fillers morphed to 5 different faces), whereas in Experiment 1 all fillers were morphed with the same face, namely, the suspect. For example, morphing faces at 60% to the suspect in Experiment 1 (highly similar condition) created fillers that were very similar to the suspect, so much so that the faces in the lineup subjectively appeared to be almost identical. By contrast, in Experiment 2, morphing different fillers to different faces rated to be dissimilar to the suspect at 60% yielded fillers that were much less similar to the suspect and quite dissimilar to each other as well (Figure 8).

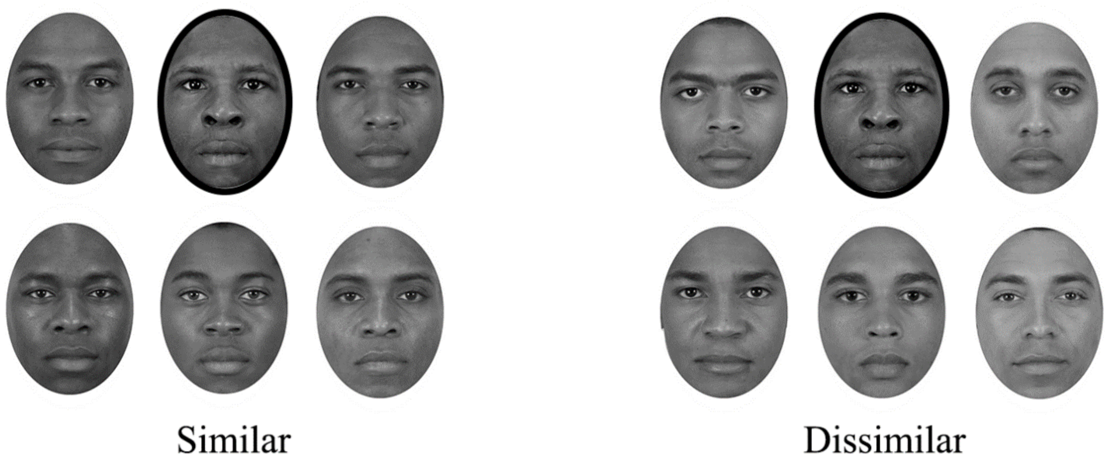

Figure 8. Examples of lineups constructed with stimuli at two similarity levels, "similar" (morphed at 60% with faces received the highest average similarity scores) and "dissimilar" (at 60% with faces received the lowest average similarity scores). Only black male stimuli are shown here. The white male stimuli were made with the same faces from Figure 5. White and black female stimuli have been made available on the OSF page.

For the main experiment, we used a 2 (filler similarity: dissimilar, similar) × 2 (target-present vs. target-absent lineups) × 2 (2-person lineup/6-person lineup) mixed factorial design. Filler similarity was a between-subject factor, while the other two were within subject factors. All photo stimuli were selected from the Chicago Face Database (CFD; Ma, Correll, & Wittenbrink, 2015).

Procedure

Each participant in the main experiment received four trials, two of which involved photos of females (black faces for one and white faces for the other) and two of which involved photos of males (black faces for one and white faces for the other). Every trial included a study phase, a 60-second distractor task and a test phase. During each study phase, the participant viewed one photo for 3 seconds. After viewing the photo, the participant was given a distractor task: playing one of the two mini games, "Tetris" or "2048", for 60 seconds. The participant then viewed either a 2-person photo lineup or a 6-person (two-row by three-column) photo lineup depending on the assigned condition, each with a 'Not Present' option underneath it. The spatial location of each photograph was randomized. The participant was given the instruction “Please choose the face you saw. If you do not recognize any of the faces, click on the ‘Not Present’ option.” On the same screen, participants were asked to assess how confident they were about their decision using an 11-point scale, ranging from 0 (not certain at all) to 10 (absolutely certain). After all four 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.

Experiment 2 was preregistered (link: https://aspredicted.org/i4kj2.pdf). The materials used in this experiment and the data reported next are available at https://osf.io/wv6tz/ (Shen & Wixted, 2023).

Results and Discussion

The proportions and frequency counts of response outcomes (suspect ID, filler ID, no ID) for TP and TA lineups across two different lineup sizes and two similarity levels were calculated and shown in Table 7.

Table 7. Frequency counts (top) and proportions (bottom) of Suspect IDs, Filler IDs, and No IDs in the 2-person or 6 person conditions for Target-Present (TP) and Target-Absent (TA) Lineups with similar and dissimilar similarity fillers.

		TP counts			TA counts		
Similarity	Size	Suspect ID	Filler ID	No ID	Suspect ID	Filler ID	No ID
Similar	2	371	41	445	81	95	681
	6	274	147	436	44	231	582
Dissimilar	2	391	27	473	76	100	715
	6	382	82	427	47	256	588

		TP proportions			TA proportions		
Similarity	Size	Suspect ID	Filler ID	No ID	Suspect ID	Filler ID	No ID
Similar	2	0.43	0.05	0.52	0.09	0.11	0.79
	6	0.32	0.17	0.51	0.05	0.27	0.68
Dissimilar	2	0.44	0.03	0.53	0.09	0.11	0.80
	6	0.43	0.09	0.48	0.05	0.29	0.66

Figure 9 presents the corresponding ROC data for Experiment 2. In the Similar condition, discriminability was scarcely affected by lineup size. This is perhaps not surprising given that filler similarity in this condition was comparable to that of the Similar condition in Experiment 1. In both cases, filler similarity deviated from the level of similarity associated with description-matched lineups by a relatively small degree (albeit in different directions). However, in the Dissimilar condition, 6-person lineups now yielded higher discriminability than 2-person lineups, which is opposite to the trend observed in Experiment 1.

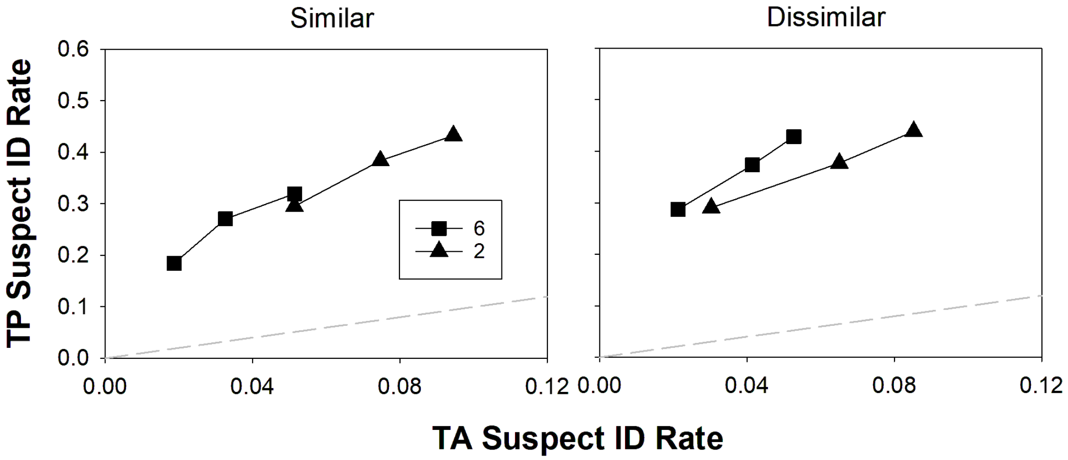

Figure 9. ROC data from the similar and dissimilar conditions of Experiment 2.

To test the significance of these results, we again fit the Independent Observations model to the data using maximum likelihood estimation. Initially, the model was fit to the data from all four conditions simultaneously, with the free parameters either fixed or free to vary across conditions in a manner consistent with what the model predicts. That is, μ_I was set to 0 and μ_G was free to vary but was constrained to be equal across both filler similarity conditions and both lineup size conditions (one parameter). In addition, as before, μ_{FTP} and μ_{FTA} were allowed to vary as a function of filler similarity (but not lineup size), adding four additional parameters. The correlation parameter was set to 0 because, once again, allowing it to vary did not significantly (or even appreciably) improve the fit.

Because there are 48 degrees of freedom in the data and 19 free parameters in the model fit, it leaves 48 – 17 = 31 remaining degrees of freedom. Thus, the expected chi square goodness-of-fit given the right model would be 31. The actual fit was $\chi^2(31) = 54.4, p = .006$, which means that the data deviated significantly from the predictions of the model. The results are shown in Table 8.

Table 8. Discriminability measures ($d'_{IG}, d'_{TP}, d'_{TA}$) and parameter estimates ($\mu_G, \mu_{FTP}, \mu_{FTA}, c_1, c_2, c_3$) based on a fit of the Independent Observations model to the data from Experiment 2.

Similarity Condition	k	d'_{IG}	d'_{TP}	d'_{TA}	μ_G	μ_{FTP}	μ_{FTA}	c_1	c_2	c_3
Similar	2	1.23	1.31	-0.10	1.23	-0.08	0.10	1.34	1.49	1.72
	6	1.23	1.31	-0.10	1.23	-0.08	0.10	1.61	1.80	2.09
Dissimilar	2	1.23	1.67	0.00	1.23	-0.44	0.00	1.31	1.47	1.75
	6	1.23	1.67	0.00	1.23	-0.44	0.00	1.46	1.61	1.89

We next asked if allowing μ_G to vary across filler similarity conditions would improve the fit even though the Independent Observations model predicts that it should not. Once again, the fit was significantly improved, $\chi^2(1) = 4.7, p = .029$, with a higher μ_G (and, therefore, higher d'_{IG}) in the dissimilar condition (Table 9). This is consistent with prior work reporting that maximizing filler dissimilarity enhanced discriminability (Colloff et al., 2021; Shen et al., 2023), and it is a result that is predicted by the Ensemble model.

Table 9. Discriminability measures and maximum likelihood parameter estimates based on a fit of the Independent Observations model to the data from Experiment 2, now with μ_G (and therefore d'_{IG}) free to vary as a function of filler similarity.

Similarity Condition	k	d'_{IG}	d'_{TP}	d'_{TA}	μ_G	μ_{FTP}	μ_{FTA}	c_1	c_2	c_3
Similar	2	1.15	1.28	-0.04	1.15	-0.14	0.04	1.28	1.43	1.66
	6	1.15	1.28	-0.04	1.15	-0.14	0.04	1.55	1.74	2.03
Dissimilar	2	1.31	1.70	-0.05	1.31	-0.39	0.05	1.36	1.53	1.81
	6	1.31	1.70	-0.05	1.31	-0.39	0.05	1.51	1.67	1.95

Finally, we asked if μ_G was additionally affected by lineup size even though there is no reason why it should be from the perspective of the Independent Observations model. By contrast, the Ensemble model predicts that increasing lineup size should enhance discriminability. Allowing μ_G to vary as a function of lineup size (2 vs. 6), separately for the Dissimilar and Similar conditions (adding two free parameters) dramatically improved the fit, $\chi^2(2) = 15.9, p < .001$ (model fits of the ROC data are presented in the Appendix). Note that this effect was almost entirely due to the effect of lineup size on μ_G in the Dissimilar condition (Table 10).

Table 10. Discriminability and correlation estimates based on a fit of the Independent Observations model to the data from Experiment 1, now with μ_G (and, therefore, d'_{IG}) free to vary as a function of both filler similarity and lineup size.

Similarity Condition	k	d'_{IG}	d'_{TP}	d'_{TA}	μ_G	μ_{FTP}	μ_{FTA}	c_1	c_2	c_3
Similar	2	1.13	1.26	-0.05	1.13	-0.13	0.05	1.28	1.43	1.66
	6	1.16	1.29	-0.05	1.16	-0.13	0.05	1.56	1.75	2.04
Dissimilar	2	1.18	1.54	-0.08	1.18	-0.35	0.08	1.32	1.48	1.76
	6	1.45	1.80	-0.08	1.45	-0.35	0.08	1.57	1.72	2.01

When the Ensemble model was fit to the same data, the parameter trends for d'_{IG} were the same as those shown in Table 10 (d'_{IG} = 0.90 and 1.37 for lineup sizes of 2 and 6, respectively, in the dissimilar condition), and goodness-of-fit for the 6-person lineups in the two filler-similarity conditions was comparable to that of the Independent Observations model. However, goodness-of-fit for the corresponding 2-person lineups was dramatically worse. As noted earlier, the likelihood functions for the Ensemble model involve an approximation that, according to the central limit theorem, becomes more accurate as lineup size increases. A lineup size of 6 is large enough to allow for a very close approximation (so fitting this model to data from 6-person lineups makes sense), but a lineup size of 2 is not large enough. Thus, fitting the Ensemble model to the 2-person lineup conditions is not entirely appropriate. Even so, when it is fit to the data, it yields the same story as that yielded by the generic signal detection model we used instead.

Finally, Figure 10 shows the detection ROCs for Experiment 2. Recall that the Independent Observations model predicts that the ROC should increase with increasing filler similarity, whereas the Ensemble model predicts the opposite trend. The effects were not as strong as in Experiment 1, but, at least for the 6-person lineup, the trend favors the Ensemble model.

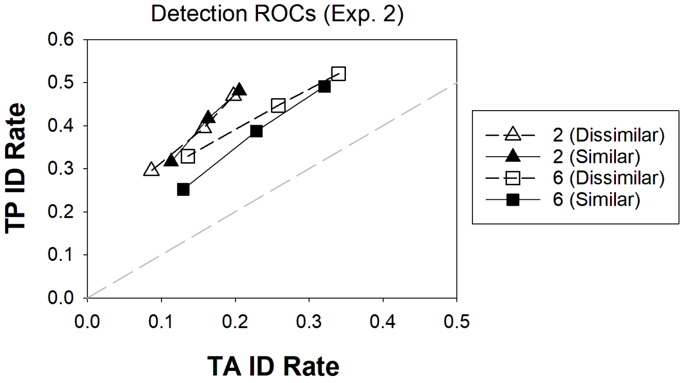

Figure 10. Detection ROC data from the similar and highly similar lineup conditions of Experiment 2. The TP ID rate is the proportion of TP lineups that resulted in a positive ID of a suspect or filler, and the TA ID rate is the proportion of TA lineups that resulted in a positive ID of a suspect or filler. The dashed line represents chance performance.

General Discussion

Previous research investigating how filler similarity affects the ability to discriminate innocent from guilty suspects in lineups supported the Ensemble model. More specifically, using fillers matched to the basic physical characteristics of the suspect (race, gender, and age), fillers who were otherwise dissimilar to the suspect enhanced discriminability (e.g., Colloff et al., 2021; Shen et al., 2023; Wells et al., 1993). However, previous research investigating how lineup size affects the ability to discriminate innocent from guilty suspects in lineups supported the Independent Observations model. More specifically, discriminability did not increase as the number of fillers in the lineup increased (Akan et al., 2020; Wooten et al., 2020), as predicted by the Ensemble model, but instead remained constant, as predicted by the Independent Observations model. The experiments reported here simultaneously manipulated both variables (filler similarity and lineup size) and found a pattern of results that is not predicted by either model: increasing lineup size impaired discriminability when highly similar fillers were used but enhanced discriminability when highly dissimilar fillers were used.

Although the observed interaction between filler similarity and lineup size is novel, the main effects we observed are consistent with prior research. For example, we observed no main effect of lineup size on discriminability under conditions that were comparable to prior studies. This was true of the Similar condition in both Experiment 1 and Experiment 2, which essentially involved description-matched fillers. That is, in both experiments, the Similar condition involved fillers who were only slightly more similar to the suspect than would be the case had we simply used description-matched fillers. Thus, these conditions were comparable to prior investigations of the effect of lineup size on discriminability, both of which used description-matched fillers (Akan et al., 2020; Wooten et al., 2020). Our results were comparable as well in that discriminability was not appreciably affected by lineup size. It was only when we deviated more substantially from the standard description-matched scenario that increasing lineup sized affected discriminability and in different directions depending on whether the fillers were highly similar or highly dissimilar to the suspect.

Our results are also consistent with prior research in which a main effect of filler similarity on discriminability was observed. In prior studies that used procedures comparable to those used here, filler similarity to the suspect was manipulated bidirectionally relative to standard description-matched fillers (e.g., Colloff et al., 2021; Shen et al., 2023; Wells et al., 1993). In those studies, increasing filler similarity to the suspect impaired discriminability (as we also observed in Experiment 1), whereas decreasing filler similarity to the suspect enhanced discriminability (as we also observed in Experiment 2). In other comparable prior work, filler similarity to the suspect was manipulated unidirectionally relative to standard description-matched fillers by making the fillers more similar to the suspect than would otherwise be the case (as we did in Experiment 1 here). The general finding from these studies is that the more similar the fillers were to the suspect, the worse the performance was (Carlson et al., 2019; Fitzgerald et al., 2015; Oriet & Fitzgerald, 2018). Again, these findings accord with what we found here in Experiment 1.

Importantly, in all of these studies, filler similarity was manipulated with respect to the *suspect* in the lineup (i.e., to the guilty suspect in TP lineups and to the innocent suspect in TA lineups). This is something that the police can do in real-world lineups. In other filler similarity studies that are less comparable to the present research, similarity was manipulated relative to the *perpetrator* in both TA and TP lineups. This is something the police cannot do because, at the time a lineup is administered, the police do not know who the perpetrator is. They only know who the suspect is. Perhaps not surprisingly, these studies generally yield a different pattern of results (e.g., Colloff et al., 2021, Experiment 2; Fitzgerald et al., 2013; Lucas, Brewer, & Palmer, 2020). For example, using this approach, Colloff et al. (2021, Experiment 2) found that the ability to discriminate innocent from guilty suspects increased as filler similarity to the perpetrator increased (the opposite of what was observed here in Experiment 1). Similarly, Lucas et al. (2020) used the same approach, and the same pattern is evident in the hit and false alarm rates associated with suspect identifications in the ROC data depicted in their Figure 2.[4] The opposite filler-similarity trends observed in these experiments are not fundamentally inconsistent with the results reported here because the procedure used in those experiments is fundamentally different from the procedure used here. Thus, different theoretical considerations apply.

The experiments reported here extended prior research by simultaneously manipulating filler similarity and lineup size. Doing so yielded a pattern of results that has not been previously observed and is not predicted by either the Independent Observations model or the Ensemble model. Specifically, increasing lineup size reduced discriminability when highly similar fillers were used (Experiment 1), but it enhanced discriminability when highly dissimilar fillers were used (Experiment 2). These findings suggest that there must be another factor affecting the ability to discriminate innocent from guilty suspects that is not included in either of the two competing models.

What might that factor be? Visual search experiments sometimes involve tasks that are analogous to lineups in that a target may or may not be presented in an array of distractors. In

[4] They also created a full ROC by cumulatively plotting the rates of all lineup decisions across confidence in target-present vs. target-absent lineups (suspect IDs, filler IDs, and lineup rejections). However, this is conceptualized as an ROC of police investigator arrest decisions (not eyewitness identification decisions), and it is not tethered to any formal model. Even if it were, it would be a model of evidence values in the brains of police investigators, not a model of memory signals in the brains of eyewitnesses, which is our exclusive focus.

that literature, many findings have pointed to the automatic ensemble perception of summary statistics, such as the average size of the items in the search set (e.g., Ariely, 2001; Chong & Treisman, 2003). However, a puzzle is that the precision of the average often does not increase with the number of items in the search set (e.g., Franconeri, Alvarez, & Enns, 2007). Why not?

One possibility, noted by Alvarez (2011), is that as the number of items increases, each item receives less attention, reducing the precision with which each item is represented. Similarly, 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). In this regard, Mazyar, van den Berg, Seilheimer, and Ma (2013) quoted Scottish philosopher Sir William Hamilton, who once noted 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).

Applied to lineups, the basic idea would be that the more fillers there are in the array, the larger σ will be, a factor not included in the equations for the Ensemble model or the Independent Observations 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. (2012, 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.

If such a noise factor is relevant to recognition memory tested using a lineup, what might it reflect, and why would it be larger when fillers are similar to the suspect (and to each other)? One possibility might be the frequency of eye fixations on the faces in the lineup. Prior eye tracking research by Flowe and Cottrell (2011) identified several variables that influence the frequency of return visits to a face in a lineup (e.g., the frequency is higher for incorrect positive identifications compared to correct positive identifications). Conceivably, when the fillers are highly similar to the suspect, participants do not scan each face only one time before making a decision but instead revisit the faces in the lineup multiple times before deciding whether or not to identify a face. Each time the memory of a face is assessed, it may slightly perturb the memory representation of that face in a nonsystematic (noisy) way. This would have the effect of reducing discriminability as lineup size increases, and repeatedly assessing the memory of a face in a lineup would presumably occur more often when the faces are similar.

A related point has been proposed in a different line of research investigating “belief bias.” In the context of the belief bias paradigm, Stephens, Dunn, and Hayes (2019) proposed that when subjective evidence is assessed only once, a decision is made based on the decision variable (x) exceeding a decision criterion (c). However, a confidence judgment is based on x', a noisy memory trace of argument strength, x. That is, $x' = x + \epsilon$, where ε is an error term. Here, we make a similar suggestion. Specifically, participants scan the faces in a lineup multiple times before making a decision, and each time they do, they update the strength of memory trace with the error component. The number of times faces are scanned before making a decision increases as filler similarity increases. If so, then the amount of noise added per additional filler in the lineup would increase as filler similarity increases.

This is a speculative interpretation, but it is testable. The first step in testing it might be to conduct an eye-tracking study like the one by Flowe and Cottrell (2011), with filler similarity manipulated across conditions. If the number of fixations increases as filler similarity increases (as predicted), the next step would be to experimentally eliminate that effect to see if now, discriminability increases with lineup size (as predicted by the Ensemble model) even when high-similarity fillers are used. One way to eliminate the effect might be to require speeded responding.

Although it has not yet been tested, if this explanation is accurate, then the overall set of results reported here would be more consistent with the Ensemble model than the Independent Observations model. When highly similar fillers are used, the noise mechanisms discussed above would explain why the results deviate from predictions made by both the Independent Observations model and the Ensemble model. However, when highly dissimilar fillers are used (theoretically minimizing the noise factor), increasing lineup size increased discriminability, as uniquely predicted by the Ensemble model.[5] The Independent Observations model has no mechanism to predict this effect. The Independent Observations model also has no mechanism to predict the effect of filler similarity on d'_{IG}, whereas the Ensemble model makes the correct prediction a priori.

Enhancing our theoretical understanding of filler similarity and lineup size seems like an important research priority given its applied relevance. For many years, the factors that reduced the information value of eyewitness decisions from lineups were social in nature. When those factors are effectively addressed, as they often are nowadays, the cognitive psychology of lineup

[5] Shen et al. (2023) reported model-recovery simulations and found that when simulated data are generated by the best-fitting Ensemble model and then fit by the Independent Observations model, allowing μ_G to vary as a function of filler similarity significantly improved the fit.

memory takes center stage. The findings reported here are presented with that relatively new state of affairs in mind.

References

Akan, M., Robinson, M., Mickes, L. B., Wixted, J., & Benjamin, A. (2021). The effect of lineup size on eyewitness identification. *Journal of Experimental Psychology: Applied, 27*(*2*), 369–392.

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(11)*, 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.

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

Chong, S. C., & Treisman, A. 2005b. Statistical processing: computing the average size in perceptual groups. *Vision Research, 45*, 891–900.

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)*, e2017292118; https://doi.org/10.1073/pnas.2017292118

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., Oriet, C., & Charman, S. D. (2013). The effect of suspect-filler similarity on eyewitness identification decisions: A meta-analysis. *Psychology, Public Policy, and Law, 19*, 151–164. https://doi.org/10.1037/a0030618

Flowe, H., & Cottrell, G. W. (2011). An examination of simultaneous lineup identification decision processes using eye tracking. *Applied Cognitive Psychology, 25*(3), 443–451. https://doi.org/10.1002/acp.1711

Franconeri, S. L., Alvarez, G. A., & Enns, J. T. (2007). How many locations can be selected at once? *Journal of Experimental Psychology: Human Perception and Performance, 33(5)*, 1003–1012.

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

Lindsay, R. C. L., Kalmet, N., Leung, J., Bertrand, M. I., Sauer, J. D., & Sauerland, M. (2013). Confidence and accuracy of lineup selections and rejections: Postdicting rejection accuracy with confidence. *Journal of Applied Research in Memory and Cognition, 2(3)*, 179–184.

Lucas, C. A., Brewer, N., & Palmer, M. A. (2021). Eyewitness identification: The complex issue of suspect-filler similarity. Psychology, Public Policy, and Law, 27(2), 151.

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, D. S., Correll, J., & Wittenbrink, B. (2015). The Chicago face database: A free stimulus set of faces and norming data. *Behavior research methods, 47(4)*, 1122-1135.

Mazyar, H., Van den Berg, R., & Ma, W.J. (2012). Does precision decrease with set size? *Journal of Vision, 12*(*6*), 10-10.

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)*, 8-8.

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

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.

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.

Shen, K. J., Colloff, M. F., Vul, E., Wilson, B. M., & Wixted, J. T. (2023). Modeling face similarity in police lineups. *Psychological Review, 130*(2), 432–461.

Shen, K. J., & Wixted, J. (2023, June 7). The Effects of Filler Similarity and Lineup Size on Eyewitness Identification. Retrieved from osf.io/wv6tz

Stephens, R. G., Dunn, J. C., & Hayes, B. K. (2019). Belief bias is response bias: Evidence from a two-step signal detection model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(2), 320–332.

Weber, N., & Brewer, N. (2006). Positive versus negative face recognition decisions: Confidence, accuracy, and response latency. *Applied Cognitive Psychology, 20*,17–31.

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.

Wells, G. L., Small, M., Penrod, S., Malpass, R. S., Fulero, S. M., & Brimacombe, C. A. E. (1998). Eyewitness identification procedures: Recommendations for lineups and photospreads.*Law and Human Behavior, 22*(6), 603–647

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. & 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(3)*, 590-604.

Yilmaz, A. S., Lebensfeld, T. C., & Wilson, B. M. (2022). The reveal procedure: A way to enhance evidence of innocence from police lineups. *Law and Human Behavior, 46*(2), 164–173.

Appendix

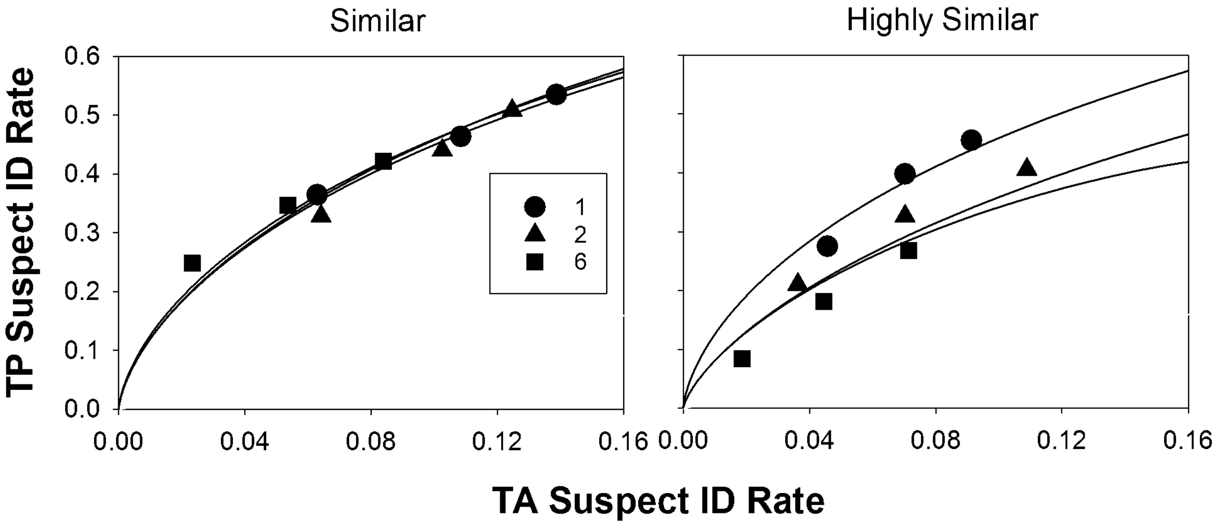

Figure A1. ROC data from Figure 5 of Experiment 1 except that the smooth curves drawn through the data were generated by the best-fitting version of the Independent Observations model (using the parameter estimates presented in Table 6).

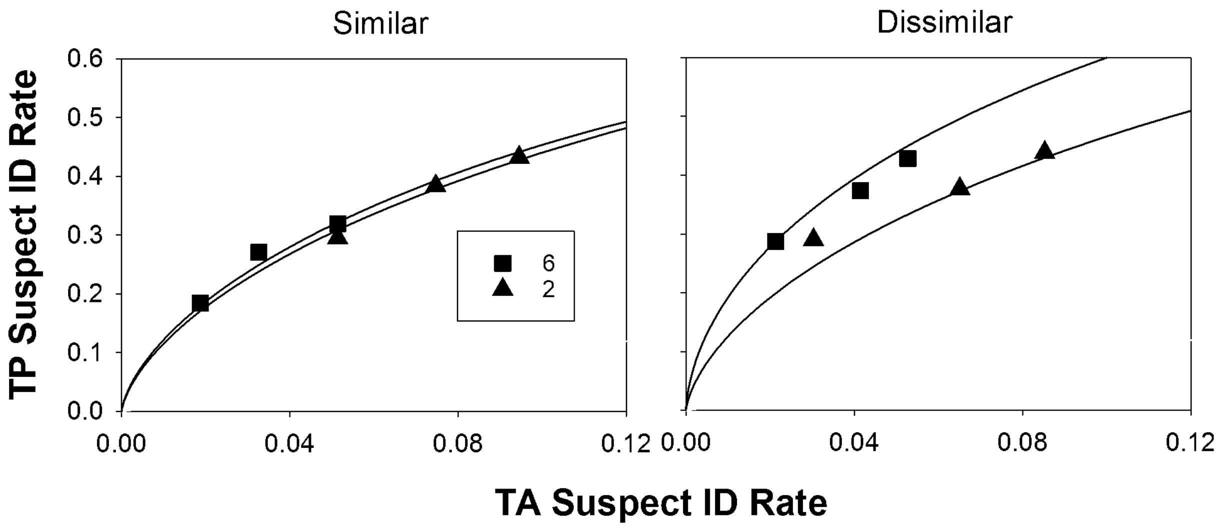

Figure A2. ROC data from Figure 9 of Experiment 2 except that the smooth curves drawn through the data were generated by the best-fitting version of the Independent Observations model (using the parameter estimates presented in Table 10).