Memory & Cognition
https://doi.org/10.3758/s13421-024-01649-x

The effect of lineup size on discriminability is dependent on filler similarity and independent of encoding strength

Allan L. Lam[1] · John T. Wixted[1]

Accepted: 25 September 2024

Abstract

A photo lineup, which consists of one suspect and several physically similar fillers, is often used by the police to test an eyewitness's memory. To optimize memory performance, how similar should the fillers be to the suspect, and how many fillers should be included in the lineup? Recent work suggests that using fillers who match the basic characteristics of the perpetrator (e.g., same age, race, and gender) but who are otherwise maximally *dissimilar* to the suspect optimizes discriminability. However, the optimal lineup size has been found to vary with filler similarity, with larger lineup sizes increasing or decreasing discriminability depending on whether low-similarity or high-similarity fillers were used, respectively. Because manipulating filler similarity at retrieval affects overall performance, here we investigated whether encoding manipulations that affect overall performance also affect how lineup size influences discriminability. In three experiments, we first replicated prior findings ($N = 502$), then reduced encoding strength by making study images blurry when low-similarity fillers were used ($N = 553$), and finally increased encoding strength by repeating study images when high-similarity fillers were used ($N = 501$). We found that whether overall performance was low or high due these encoding manipulations, discriminability still increased as a function of lineup size when low-similarity fillers were used and decreased as a function of lineup size when high-similarity fillers were used. Thus, lineup size has opposing effects on discriminability when task difficulty is manipulated at retrieval, which narrows the theoretical explanations for why that effect is observed.

Keywords Eyewitness discriminability · Filler similarity · Lineup size · Signal detection theory

Introduction

Lineups are frequently used in the criminal justice system to assess eyewitness memory. A lineup consists of one suspect (who is innocent or guilty) and a number of physically similar fillers who are all known to be innocent. A witness can identify the suspect, identify a filler (a known error), or reject the lineup. Given the significance of suspect identifications at criminal trials, a great deal of eyewitness identification research has been conducted to understand how different variables affect lineup performance. *System variables* are especially important in this domain because they are variables that can be manipulated by law enforcement agencies during the investigation (Wells, 1978). One key system variable is the number of fillers in a lineup (*lineup size*), and another key system variable is how similar the fillers are to the suspect (*filler similarity*). Two longstanding recommendations are that a lineup should have at least five fillers and that the fillers should be sufficiently similar to the suspect that the suspect does not stand out (Wells et al., 1998, 2020). Although both lineup size and filler similarity are important to consider in creating a fair lineup, their optimal levels are yet to be fully understood. To optimize lineup performance, how large should the lineup be, exactly, and how similar should the fillers be to the suspect, exactly?

Recent laboratory-based lineup research that attempted to identify the optimal levels of these two variables defined optimality in terms of maximizing an eyewitness's ability to discriminate innocent from guilty suspects (*discriminability*). This focus on discriminability differs from the focus of earlier efforts to optimize filler similarity (e.g., Fitzgerald et al., 2015; Luus & Wells, 1991; Oriet & Fitzgerald, 2018). Although it is not the only possible definition of optimality,

✉ Allan L. Lam
alam@ucsd.edu

✉ John T. Wixted
jwixted@ucsd.edu

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

Published online: 07 November 2024

it is a reasonable definition because a procedure with higher discriminability can be used to achieve both a higher hit rate and a lower false alarm rate than a procedure with lower discriminability. The hit rate is the proportion of target-present (TP) lineups resulting in a correct identification of the guilty suspect, and the false alarm rate is the proportion of target-absent (TA) lineups resulting in an incorrect identification of the innocent suspect.

How is discriminability measured when a lineup is used? When confidence ratings are obtained for lineup decisions, they can be used to compute multiple hit and false alarm rates. Discriminability can then be empirically assessed by measuring the area under a confidence-based receiver operating characteristic (ROC). Alternatively, in terms of underlying memory signals, discriminability can be measured using d'. In a standard list memory paradigm, d' represents the degree to which the distribution of memory signals generated by previously seen targets overlaps with the distribution of memory signals generated by new foils. For a lineup task, this measure (now denoted d'_{IG}) represents the degree to which the distribution of memory signals generated by innocent suspects in TA lineups overlaps with the distribution of memory signals generated by guilty suspects in TP lineups. As described in the Appendix (Online Supplemental Material), it can be estimated by fitting a simple signal detection model to the full set of lineup data obtained for a given condition (i.e., suspect IDs, filler IDs, and lineup rejections).

Our main focus is on d'_{IG}. More specifically, how can filler similarity and lineup size be jointly manipulated in such a way as to maximize d'_{IG}? With regard to filler similarity, an initial and basic consideration is that every filler should match the witness's description of the perpetrator (Wells et al., 1993). The reason is that it would make little sense to include fillers who do not possess features of the perpetrator known to be represented in the witness's memory. In a laboratory study, the use of description-matched fillers can be approximated by ensuring that the fillers in the lineup match the basic physical characteristics of a previously studied face (e.g., age, race, gender, and hair color/style). Yet, in a pool of potential fillers who match these basic physical characteristics, there will still be considerable variability in how similar they are to the suspect in the lineup. A typical description-matched filler would have some baseline level of similarity, but less typical description-matched fillers would have either more or less similarity to the suspect than that. Thus, the question arises as to just how similar to the suspect the fillers that are drawn from that pool should be in order to maximize d'_{IG}.

Colloff et al. (2021) manipulated filler similarity relative to the innocent in TA lineups and to the guilty suspect in TP lineups. Filler similarity was measured via ratings provided by an independent group of participants, and their ROC data revealed that discriminability was maximized by *minimizing* filler similarity. More specifically, the hit rate increased as filler similarity decreased, but the false alarm rate remained constant. The constant false alarm rate as a function of filler similarity may not be an intuitive result, but it simply means that making a description-matched filler less similar to the innocent suspect in a TA lineup (a suspect who does not strongly match the witness's memory of the perpetrator) does not make that suspect stand out as differentially matching the witness's memory of the perpetrator. In a TP lineup, by contrast, making a filler less similar to the guilty suspect *does* make the suspect differentially stand out in memory. Recently, Shen et al. (2023) replicated these findings using a facial morphing procedure to manipulate filler similarity.

With regard to lineup size, as mentioned above, the National Institute of Justice long ago recommended using at least five fillers in a lineup (National Institute of Justice, 1999), a recommendation that remains in force today (Wells et al., 2020). Nevertheless, the optimal lineup size is still an unsettled issue. Although many studies have demonstrated the diagnostic superiority of lineups over a one-person lineup known as a "*showup*" (e.g., Colloff & Wixted, 2020), subsequent research found no significant effects on discriminability beyond a lineup size of two when standard description-matched fillers are used (Akan et al., 2021; Wooten et al., 2020). It therefore seems as if optimizing lineups can be achieved by minimizing filler similarity to the suspect (from a pool of description-matched fillers) while using any lineup size of two or more. However, some recently reported results suggest that there might be more to the story because these two variables (filler similarity and lineup size) interact in unexpected ways.

Shen et al. (2024) recently reported that the effect of lineup size on discriminability varied as a function of filler similarity. In all conditions, description-matched fillers were used. These fillers will have some degree of similarity to the suspect in the lineup (innocent or guilty) because they will share the same race, gender, and approximate age. From this baseline level of similarity, Shen et al. (2024) manipulated filler similarity to the suspect in opposite directions (lower or higher). When low-similarity fillers were used, discriminability (measured by d'_{IG}) was relatively high, consistent with Colloff et al. (2021). In addition, increasing lineup size increased discriminability. This effect is consistent with a prediction made by a signal-detection model known as the *Ensemble model* (Wixted et al., 2018). The Ensemble model holds that identification decisions are based on how familiar a face is relative to the average familiarity of the faces in the lineup. Because the average familiarity is more precisely determined the more faces there are in the lineup, discriminability should increase with lineup size.

The Ensemble model's prediction about the beneficial effect of increasing lineup size is independent of filler similarity. Thus, the fact that the prior work mentioned above using description-matched fillers (with baseline familiarity) found no effect of lineup size on discriminability (Akan et al., 2021; Wooten et al., 2020) is contrary to the model's predictions. Moreover, Shen et al. (2024) found that when filler similarity was increased above baseline (i.e., when high-similarity description-matched fillers were used), overall performance was lower, as expected, and increasing lineup size now *decreased* discriminability. This effect is the opposite of what the Ensemble model predicts.

To account for these unexpected findings, Shen et al. suggested that as filler similarity increases, a noise factor may increasingly affect performance (Ariely, 2001; Mazyar et al., 2012, 2013). What the noise factor might be is not clear, but they suggested it might reflect repeated viewing of similar faces while trying to make an identification, with each new look slightly perturbing the memory signal in a cumulative fashion. Such back-and-forth scanning prior to making a decision would presumably increase as filler similarity increased.

An alternative possibility is that the effect has nothing to do with the effect of high-similarity fillers at retrieval but instead reflects overall task difficulty. In other words, in the high-similarity condition, the task is hard, and performance is correspondingly low; in the low-similarity condition, the task is easy, and performance is correspondingly high. Perhaps any manipulation that harms overall discriminability will be associated with a reduction in discriminability as lineup size increased. If so, then making the task easier in the high-similarity condition (e.g., by allowing extra time to study the perpetrator at encoding) will minimize the deleterious effect of increasing lineup size and might even reverse the effect. Similarly, increasing task difficulty in the low-similarity condition (e.g., by making it harder to encode the perpetrator's face) might reduce the beneficial effect of increasing lineup size and might even reverse it.

To investigate this issue, we first replicated the trends reported by Shen et al. (2024) using an expanded stimulus set. We then conducted a second lineup-size experiment using low-similarity fillers and manipulated the overall level of performance by manipulating blurriness of the study faces at the study phase. By making the target blurry enough, overall memory performance was pushed down to a level similar to what is observed when high-similarity fillers are used (despite low-similarity fillers being used at retrieval). In a third experiment using high-similarity fillers, we manipulated the overall level of performance by varying target repetition at the study phase. Repeating the study targets five times raised memory performance to a level similar to what is observed when low-similarity fillers are used (despite high-similarity fillers being used at retrieval).

Experiment 1: Replication

The first experiment replicated Shen et al. (2024) using an expanded stimulus set. Whereas Shen et al. (2024) tested the effect of lineup size using high- and low-similarity fillers in separate experiments, we tested them in a single experiment. The lineup sizes were k =1, 2, and 6.

Participants

Participants were recruited through Amazon Mechanical Turk (Mturk) with additional screening implemented using Cloud Research's MTurk toolkit to ensure data quality. We collected data from a total of 970 participants, but only those who passed various exclusion criteria were included.[1] First, we only included the 668 participants who (1) successfully passed the attention check question and (2) responded "no" when asked, "have you seen these faces before?" The attention check question was "What were you asked to remember?" and the correct answer was "Face." Another 93 were excluded because they had duplicated IP addresses, and we kept only their first entries. Another 65 participants were excluded either because they took more than 30 s or less than 0.5 s for their first click on the lineup page. Finally, eight participants were excluded because they all used the identical phrase "memory power" when asked what they think is the purpose of the study. That left 502 for final analysis (*M*age = 41.1 years).

After exclusion, the 502 participants included 44.2% male (222), 55.3% female (278), 0.2% other (1), and 0.2% prefer not to state (1), with the ethnicity distribution being: 7.4% African-American (37), 7.8% Asian (39), 1.2% Mexican–American (6), 0% Filipino (0), 4.4% Latino (22), 2.6% Native-American (13), 74.7% Caucasian (375), 1.6% Other/Undeclared (8), 0.4% Prefer not to state (2).

Design and materials

The experimental design was a 2 (filler similarity: high-similarity vs. low-similarity) × 2 (lineup type: target-present vs. target-absent lineups) × 3 (lineup size: showup vs. two-person lineup vs. six-person lineup) mixed factorial design. Filler-similarity was a between-subject factor, while lineup type and lineup size were within subject factors. Therefore, each participant completed six trials, three TP and three TA trials (one at each lineup size). Whereas Shen et al. (2024) used only White and Black American faces, we included all available male and female faces across three racial groups from the Chicago Face Database to create our stimuli pool (Ma et al., 2015). A total of 489 faces were included with the following breakdown: 93 White American males, 90 White American females, 93

[1] In a given between-subjects filler similarity condition, G*Power estimated we needed ~ 800 observations to detect a small Cramér's V effect size of .10 for the lineup size manipulation with .80 power using an alpha level of .05. Because each participant in a given filler similarity condition was tested three times, this would mean testing ~ 270 participants for each condition (3 × 270 = 810 observations). Since we had two filler similarity conditions, we needed 2 × 270 = 540 participants in all. We aimed for ~ 600 participants in each of the three experiments reported here and ended up with between 500 and 600 after exclusions.

Black American males, 104 Black American females, 52 Asian American males, and 57 Asian American females (all faces from the database fell within the 20- to 30-year-old age range). Thus, for a given lineup, the faces shared key features: race, gender, and age. In this way, we operationalized the use of description-matched fillers. The order of race and gender was counterbalanced. Within each race and gender group, the target face for TP lineups was randomly designated.

As described below, a morphing procedure was used to manipulate filler similarity. Creating high-similarity fillers was straightforward because it simply required morphing the filler faces toward the innocent or guilty suspect in the lineup. In other words, creating high-similarity fillers involves morphing all of the faces towards the same face (the suspect). Creating dissimilar fillers, by contrast, is less straightforward because it involves morphing the fillers away from suspect and from each other. To do so, and as described next, potential fillers were independently rated for how similar they were to the suspect in the lineup, and the five least similar fillers were tagged. These fillers could have been used as the fillers for the low-similarity condition, but because a morphing procedure was used to create the high-similarity fillers, we used a morphing procedure to create the low-similarity fillers as well.

Preliminary similarity rating study

Similarity ratings for White and Black American males and females were directly inherited from Shen et al.'s Experiment 2, which used faces from the same database, but the Asian American category was not included in their study. Therefore, an additional pilot study was conducted to collect similarity ratings of Asian American male and female fillers relative to the designated TP target face. For this study, we recruited 97 participants from the UC San Diego SONA system, where all participants were undergraduate students attending the institution and completed the study for class credit. Participants were randomly assigned to the Asian American male (49 participants) and Asian American female (48 participants) condition, and they were instructed to rate the similarity between each filler and TP target face on a scale of 1 (highly dissimilar) to 7 (highly similar). The five faces with the *lowest* average similarity score were tagged as the dissimilar faces that would be morphed with other faces to create the low-similarity fillers. In addition, the three most median faces (i.e., the median face and the next ranked face above it and below it) were selected as the designated TA faces.

Manipulating filler similarity

Each racial-gender category consisted of four types of photos: (A) one photo of the randomly designated target face for TP lineups, (B) three photos of the randomly designated innocent suspect face for TA lineups, (C) five photos of the rated most dissimilar faces, and (D) the remaining fillers.

Suspect similarity was manipulated by morphing the above photo types to create composites of fillers using Fantamorph face-morphing software. High-similarity fillers for TP lineups were face composite morphs consisting of 60% designated TP target face (Type A) and 40% of the remaining filler (Type D). A similar procedure was used to create high-similarity fillers between innocent suspects (Type B) and the remaining fillers (Type D).

Because morphing fillers to the suspects can only increase filler similarity to varying degrees, a different approach was needed to create low-similarity fillers. To create low-similarity fillers, all remaining fillers (Type D) were first evenly divided into five folders (Folders 1–5 are available at https://osf.io/5z4y7/), basically one folder for each face rated as being most dissimilar. Afterward, every face in each folder was morphed by pairing the dissimilar face (i.e., folder 1 with the first face in Type C, folder 2 with the second face in Type C, etc.). Low-similarity fillers were face composites morphed by 60% of the dissimilar face (Type C) and 40% of the remaining filler (Type D) in each folder. Examples are shown in Fig. 1.

There were six possible combinations of within-subject conditions across the trials: one target-present and target-absent lineup for each lineup size (namely, a showup, a two-person lineup, and a six-person lineup). These combinations were randomized, and each combination corresponded to a different racial-gender category. During the study phase, participants encoded either a target face or a non-target face. The target faces were always presented at retrieval even if the non-target face had been studied at encoding. Thus, this target face was the guilty suspect in TP lineups and the innocent suspect in TA lineups. On each subsequent test trial, the memory test included the suspect and a certain number of description-matched fillers (0 or 1, or 5). The test faces were presented in a 2 × 3 array for the six-person lineups, a 2 × 1 array for the two-person lineups, and a 1 × 1 array for the showup, with the array position of the faces randomized.

Procedure

Each participant received six trials. Each trial included a 2-s study phase, a 120-s distractor task, and a test phase. During each study phase, the participant viewed one photo for 2 s and then proceeded to the distractor task. The distractor task was to play one of the two mini games randomized in assignment, "Tetris" or "2048," for 120 s. The participant then viewed either a showup, a two-person photo lineup (two-row), or a six-person (two-row by three-column) photo lineup depending on the assigned condition. Underneath each lineup was a

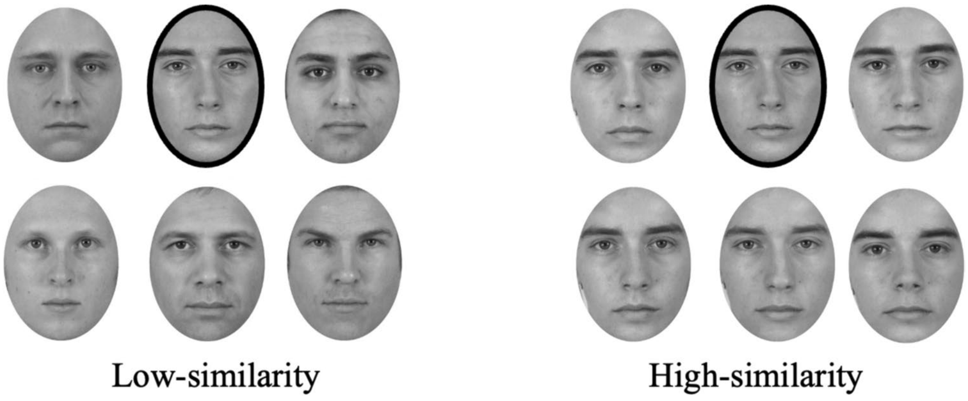

Fig. 1 Examples of lineups constructed with stimuli of "low-similarity fillers" (morphed with 60% rated dissimilar faces) and "high-similarity fillers" (morphed with 60% designated TP target face)

"Not Present" option for participants to reject the lineup if the target was not present. The array 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 using an 11-point scale, ranging from 0 (not certain at all) to 10 (absolutely certain). After completing all six trials, participants were asked about their demographic information, what they had been asked to study (the attention check question), whether they previously participated in this study, and what they thought was the purpose of the study.

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 conditions are shown in Table 1, and the corresponding ROC curves are shown in Fig. 2. Note that the showups (lineup size = 1) in the two similarity conditions did not include fillers, so there are no filler IDs for the Showup condition. Given the absence of fillers, TA and TP showups were therefore procedurally identical across suspect-filler-similarity levels. However, in one condition, the showups were presented amongst trials of lineups involving high-similarity fillers, and in the other condition, they were presented amongst trials of lineups involving low-similarity fillers. Thus, despite being procedurally identical, the data for showups are presented separately for each filler similarity condition.

The ROC curves shown in Fig. 2 revealed that, for the low-similarity fillers, discriminability increased as lineup size increased, as predicted by the Ensemble model and as observed by Shen et al. (2024). By contrast, for the high-similarity fillers, the reverse pattern was observed, with the showup now yielding the highest discriminability (with lineup sizes of 2 and 6 yielding similar levels of performance). This reversed effect was also observed by Shen et al. (2024).

As noted earlier, the designated TA suspect is conceptually equivalent to a filler, so the designated TA suspect ID rate would ideally be the same as the estimated TA suspect ID rate: (TA filler IDs + TA suspect IDs)/lineup size. Figure 3 presents the ROC data with the false alarm rate computed in this manner. For low-similarity fillers, the same pattern shown in Fig. 2 is evident again. For high-similarity fillers, discriminability for the showup condition still exceeded that of the two-person lineups but is now essentially the same as that for the six-person lineup.

We next estimated d'_{IG} in each condition by fitting a standard (Independent Observations) signal detection model to the data, with details presented in the Appendix in the OSM. The model-fitting algorithms we used are implemented in an eyewitness identification analysis toolkit (Mickes et al., 2023). The Independent Observations model assumes that the memory signals for the guilty suspect are drawn from a Gaussian distribution with mean μ_G, and the memory signals for the innocent suspect (and the innocent fillers in both TA and TP lineups) are drawn from a Gaussian distribution with mean μ_I. Because both distributions are assumed to have the same standard deviation, σ, $d'_{IG} = \frac{\mu_G - \mu_I}{\sigma}$. For simplicity, we set $\mu_I = 0$ and $\sigma = 1$, such that $d'_{IG} = \mu_G$. Thus, estimating how the parameter μ_G is affected by various experimental manipulations is tantamount to estimating how d'_{IG} varies across those conditions.

We began by constraining μ_G to be equivalent for all six conditions (two filler similarity conditions by three lineup size conditions). Allowing it to differ as a function of filler similarity significantly improved the fit, χ^2 (2) = 58.89, $p < .001$. This result indicates that d'_{IG} for the low-similarity condition (2.19) significantly exceeded that

Table 1 Frequencies (top) and proportions (bottom) of Suspect IDs, Filler IDs, and No IDs in the one-person (showup), two-person, and six-person lineup conditions for target-present (TP) and target-absent (TA) lineups with low-similarity and high-similarity 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
Low-similarity	1	234	–	17	56	–	195
	2	236	3	12	33	51	167
	6	228	17	6	16	92	143
High-similarity	1	230	–	21	40	–	211
	2	204	37	10	33	45	173
	6	160	76	15	21	87	143
		TP proportions			TA proportions		
Similarity	Size	Suspect ID	Filler ID	No ID	Suspect ID	Filler ID	No ID
Low-similarity	1	0.93	–	0.07	0.22	–	0.78
	2	0.94	0.01	0.05	0.13	0.20	0.67
	6	0.91	0.07	0.02	0.06	0.37	0.57
High-similarity	1	0.92	–	0.08	0.16	–	0.84
	2	0.81	0.15	0.04	0.13	0.18	0.69
	6	0.64	0.30	0.06	0.08	0.35	0.57

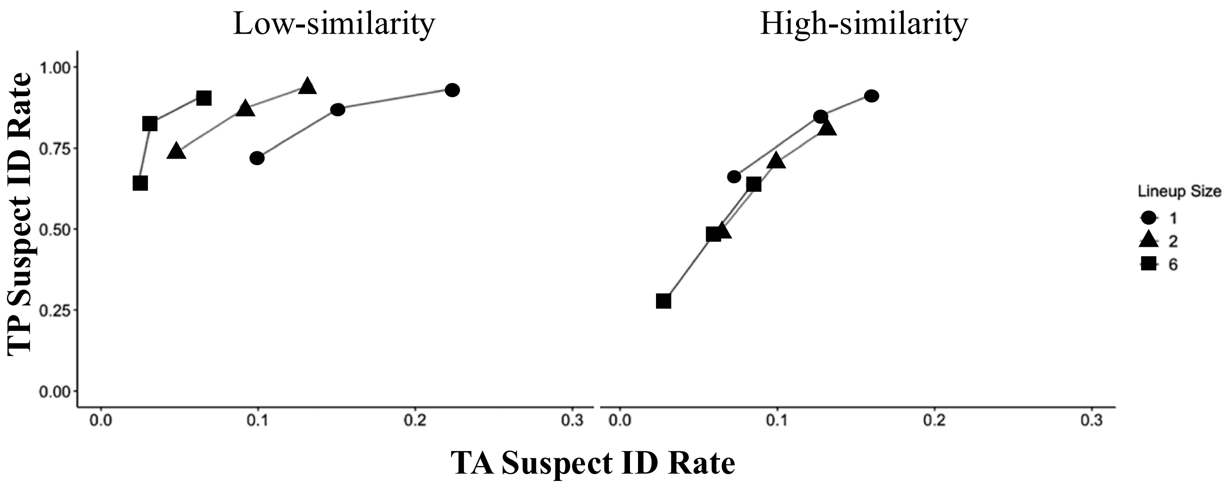

Fig. 2 Receiver operating characteristic (ROC) data from the low-similarity and high-similarity conditions of Experiment 1. The target-present (TP) suspect ID rate is the proportion of TP lineups that resulted in a suspect ID. The target-absent (TA) suspect ID rate is the proportion of designated innocent suspect chosen at lineup. Note that filler IDs from TP lineups are not represented in these ROC plots, but they are included when models are fit to the data

for the high-similarity condition (1.98). We next asked whether allowing μ_G to vary within each filler-similarity condition would significantly improve the fit. Indeed, doing so resulted in a significant improvement of the fit, $\chi^2(2) = 15.25$, $p < .001$ for the low-similarity condition and $\chi^2(2) = 7.65$, $p = .022$ for the high-similarity condition. The final estimates are shown in Table 2.

Overall, Experiment 1 replicated the results reported by Shen et al. in that the data once again indicated that filler-similarity plays a role in determining the effect of lineup size on discriminability. As mentioned above, a possible explanation is that high similarity fillers at retrieval introduced a noise factor that harmed discriminability as a function of lineup size. Shen et al. (2024) suggested that this noise factor is specifically created from highly similar fillers, thus, theoretically, this noise factor should only arise at the retrieval phase when these fillers are present. However, it might be the case that any manipulation that harms overall discriminability, including manipulations at encoding, would be associated with a reduction in

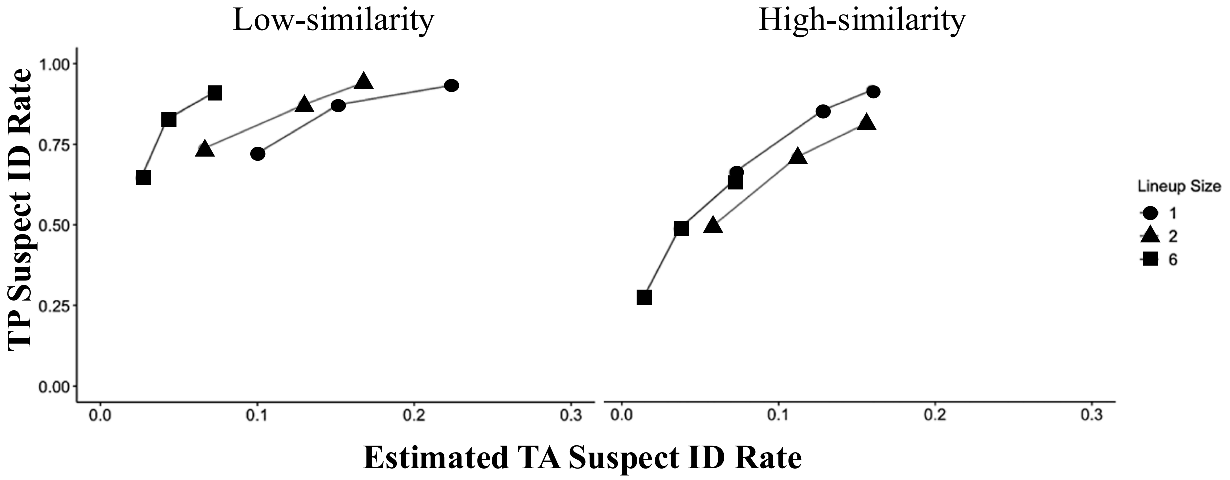

Fig. 3 Receiver operating characteristic (ROC) data from the low-similarity and high-similarity conditions of Experiment 1. The target-present (TP) suspect ID rate is the proportion of TP lineups that resulted in a suspect ID. The estimated target-absent (TA) suspect ID rate is the proportion of TA lineups that resulted in a filler ID divided by the lineup size of one or two or six (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

Table 2 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 lineup size

Similarity	Size	d'_{IG}
Low-similarity	1	2.07
	2	2.23
	6	2.27
High-similarity	1	2.15
	2	1.86
	6	1.94

discriminability as lineup size increased. We next investigated that issue for the case of low-similarity fillers.

Experiment 2: Low-similarity fillers

In Experiment 2, we manipulated overall difficulty by manipulating target blurriness at encoding across two levels (not blurry vs. blurry). As in Experiment 1, the facial stimuli at retrieval were not blurry. The lineup sizes were again k =1, 2, and 6.

Participants

Participants were again recruited through Amazon Mechanical Turk (Mturk) with Cloud Research's MTurk toolkit. In total, 553 participants (*M*age = 44.02 years) were included after passing the same deliberate sets of exclusion rules. We only included participants who successfully passed the attention check question and have not participated in the replication study, as we reused the same low-similarity fillers at lineups. This time, however, we did not exclude participants who spent more than 30 s on a given trial, as the task was purposefully designed to be more demanding when the faces were blurry at encoding. The additional question "What do you think is the purpose of the study" was still included, but there were no participants who responded with unusual comments. After exclusion, the 553 participants included 32.0% male (177), 65.8% female (364), 1.4% other (8), and 0.7% prefer not to state (4), with the ethnicity distribution being: 11.9% African-American (66), 5.2% Asian (29), 2.5% Mexican–American (14), 0.7% Filipino (4), 2.7% Latino (15), 1.4% Native-American (8), 73.2% Caucasian (405), 2.0% Other/Undeclared (11), 0.2% Prefer not to state (1).

Design and materials

This experiment used a 2 (Target Blurriness: blurred vs. non-blurred) × 2 (lineup type: target-present vs. target-absent lineups) × 3 (lineup size: showup vs. two-person lineup vs. six-person lineup) mixed factorial design. Target blurriness was a between-subject factor, while lineup type and lineup size were within-subject factors. Therefore, each participant completed six trials of lineup tasks within the study. Our stimulus pool was the same as Experiment 1. Experimental logistics, materials, and procedures were also largely consistent with Experiment 1. There were only two differences in the materials used. First, participants in the blurry condition encoded the study faces blurred at the level of 40 radial blur (see examples in Fig. 4). Second, the retrieval phase always contained low-similarity fillers. Participants assigned to the blurred condition only saw blurred faces at encoding across all six trials, and participants assigned to the non-blurred condition only saw non-blurred faces at encoding across all six trials.

Procedure

Before the experiment began, participants saw an instruction screen indicating that the faces they would see were expected to be blurry or not blurry during the study phase. Apart from that, all procedures are the same as Experiment 1.

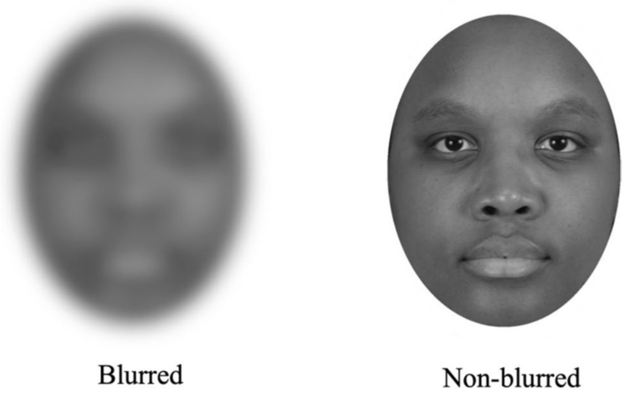

Fig. 4 Examples of non-blurred and blurred (40 radial blur) stimuli at the study

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 levels of blurriness are shown in Table 3, and the corresponding ROC curves are shown in Figs. 5 and 6. The general patterns are consistent with what would be expected from a memory-strength manipulation. For example, many recognition memory models assume that the location of the decision criterion is determined by a likelihood ratio computation, in which case a "mirror effect" should be observed (Glanzer & Adams, 1990; Glanzer et al., 1993; McClelland & Chappel, 1998; Osth et al., 2017; Semmler et al., 2018; Shiffrin & Steyvers, 1997; Stretch & Wixted, 1998; Wixted & Gaitan, 2002). The mirror effect refers to a common empirical regularity according to which the hit rate increases, and the false alarm rate decreases as overall recognition accuracy decreases. The data in Table 3 exhibit that pattern. For example, as overall accuracy decreases from the non-blurred to the blurred condition, the hit rate (i.e., suspect ID rate in TP lineups) decreases and the false alarm rate (i.e., suspect ID rate in TA lineups) increases for all three lineup size conditions. Correspondingly, the No ID rate increases in TP lineups and the No ID rate decreases in all three lineup size conditions.

The ROC curves shown in Figs. 5 and 6 indicate that in the non-blurred condition, discriminability increased with lineup size, as in the low-similarity conditions of Experiment 1 here and of Shen et al. (2024). Although overall performance was lower in the blurred condition (by design), the same trend was still observed. In other words, we did not see the reversal in

Table 3 Frequencies (top) and proportions (bottom) of Suspect IDs, Filler IDs, and No IDs in the one-person (showup), two-person, and six-person lineup conditions for target-present (TP) and target-absent (TA) lineups with non-blurred and blurred study faces. The "–" symbols represent nonexistent filler data for showups

		TP counts			TA counts		
Blurriness	Size	Suspect ID	Filler ID	No ID	Suspect ID	Filler ID	No ID
Non-blurred	1	249	–	21	31	–	239
	2	252	0	18	12	52	206
	6	244	9	17	14	93	163
Blurred	1	199	–	84	111	–	172
	2	195	23	65	64	118	101
	6	164	82	37	31	163	89
		TP proportions			TA proportions		
Blurriness	Size	Suspect ID	Filler ID	No ID	Suspect ID	Filler ID	No ID
Non-blurred	1	0.92	–	0.08	0.11	–	0.88
	2	0.93	0.00	0.07	0.04	0.19	0.76
	6	0.90	0.03	0.06	0.05	0.34	0.60
Blurred	1	0.70	–	0.30	0.39	–	0.61
	2	0.69	0.08	0.23	0.23	0.42	0.36
	6	0.58	0.29	0.13	0.11	0.58	0.31

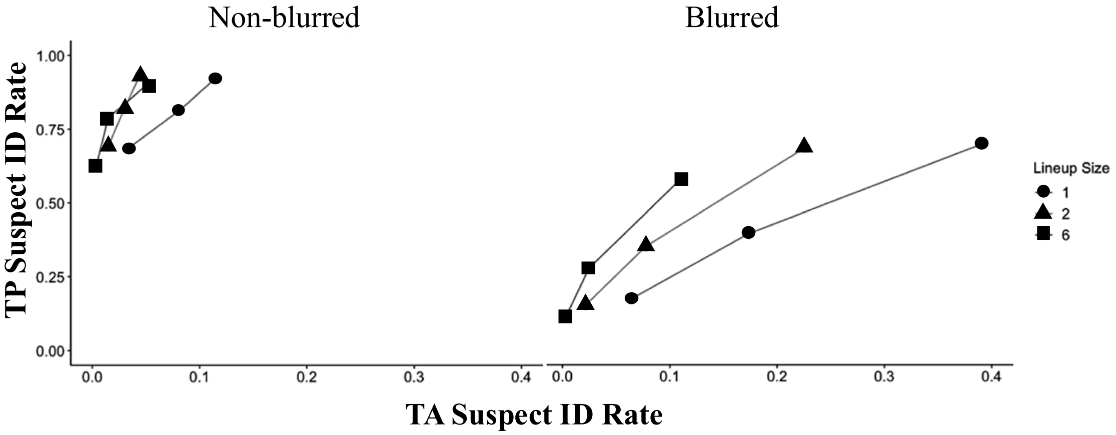

Fig. 5 Receiver operating characteristic (ROC) data from the non-blurred and blurred conditions of Experiment 2. The target-present (TP) suspect ID rate is the proportion of TP lineups that resulted in a suspect ID. The target-absent (TA) suspect ID rate is the proportion of designated innocent suspect chosen at lineup. Note that filler IDs from TP lineups are not represented in these ROC plots, but they are included when models are fit to the data

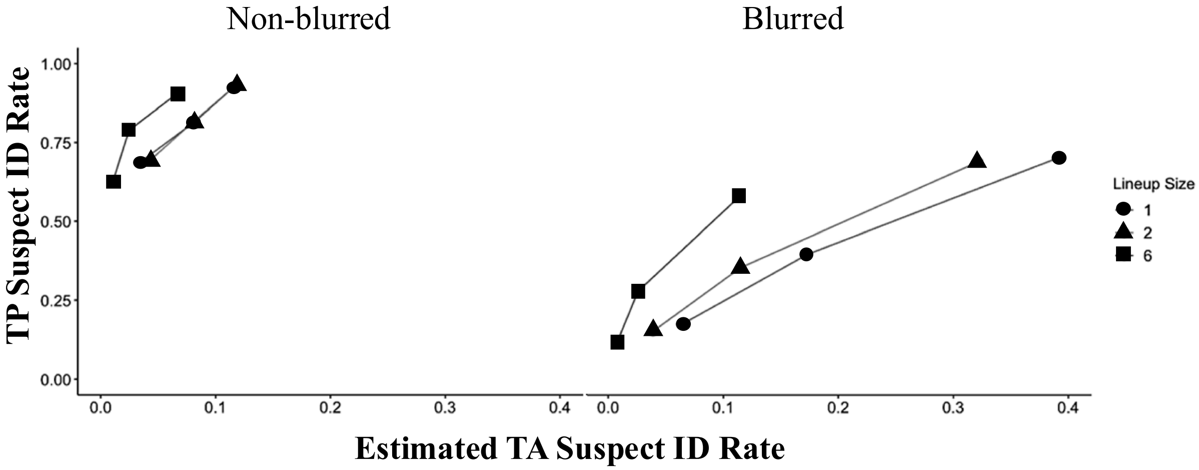

Fig. 6 Receiver operating characteristic (ROC) data from the non-blurred and blurred conditions of Experiment 2. The target-present (TP) suspect ID rate is the proportion of TP lineups that resulted in a suspect ID. The estimated target-absent (TA) suspect ID rate is the proportion of TA lineups that resulted in a filler ID divided by the lineup size of 1 or 2 or 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 effect of lineup size that occurs when overall performance is reduced by the use of high-similarity fillers.

We next estimated d'_{IG} in each condition by fitting a signal detection model, with details again presented in the Appendix (OSM). We began by constraining d'_{IG} to be equivalent for all six conditions (two blurriness conditions by three lineup size conditions). Allowing it to differ as a function of blurriness condition significantly improved the fit, $\chi^2(2) = 275.93$, $p < .001$. This result indicates that d'_{IG} for the non-blurred condition (2.57) significantly exceeded that for the blurred condition (1.02). We next asked whether allowing d'_{IG} to vary within each blurriness condition would significantly improve the fit. Indeed, doing so resulted in a significant improvement of the fit, $\chi^2(2) = 27.39$, $p < .001$ for the non-blurred condition and $\chi^2(2) = 37.17$, $p < .001$ for the blurred condition. For both conditions, discriminability was highest in the lineup size 6 condition (Table 4).

Experiment 3

The beneficial effects of increasing lineup size persisted when low-similarity fillers were used even when lowering discriminability to a level similar to that observed when high-similarity fillers are used. A parallel question remains:

Table 4 Discriminability estimates from Experiment 2, with d'_{IG} free to vary as a function of lineup size

Condition	Lineup size	d'_{IG}
Non-blurred	1	2.46
	2	2.70
	6	3.08
Blurred	1	0.73
	2	0.99
	6	1.41

when using high-similarity fillers, will the deleterious lineup size effect persist if we increase discriminability to approximately the same level observed when low-similarity fillers are used? To investigate this issue, we used high-similarity fillers and made the task easier by presenting the target face five times during the study phase (as opposed to only one time in the control condition). Note that the low repetition condition (one-time) is identical to the high-similarity condition from the first experiment. The lineup sizes were still k =1, 2, and 6.

Participants

Participants were recruited through Amazon Mechanical Turk (Mturk) with Cloud Research's MTurk toolkit. In total, 501 participants (*M*age = 43.1 years) were included after passing the same deliberate sets of exclusion rules. We only included participants who successfully passed the attention check question and had not participated in Experiments 1 and 2, as we reused the same high-similarity fillers at lineups. We excluded participants who spent less than 0.5 s and more than 30 s at a given trial because the task was not as demanding as in Experiment 2. All participants responded "What do you think is the purpose of the study" with reasonable answers. After exclusion, the 501 participants included 30.7% male (154), 67.7% female (339), 1.6% other (8), and 0% prefer not to state (0), with the ethnicity distribution being: 7.8% African-American (39), 3.8% Asian (19), 2.0% Mexican–American (10), 0.4% Filipino (2), 4.8% Latino (24), 1.4% Native-American (7), 77.4% Caucasian (388), 2.2% Other/Undeclared (11), 0.2% Prefer not to state (1).

Design and materials

This experiment used a 2 (Target Repetition: Repeated vs. Unrepeated) × 2 (lineup type: target-present vs. target-absent lineups) × 3 (lineup size: showup vs. two-person lineup vs. six-person lineup) mixed factorial design. Target repetition was a between-subject factor, while lineup type and lineup size were within-subject factors. Therefore, each participant completed six trials. We used the same racial-gender categories and the same faces from the stimuli pool as the previous studies. Experimental logistics, materials, and procedures were also largely consistent with Experiments 1 and 2 with two differences: First, participants in the repeated condition were exposed to the 2-s study phase five times. The target face was presented for 2 s within each repetition, along with a message saying "You will see the face again next. Please try your best to remember the face" between each repetition. In the unrepeated condition, participants simply studied

Table 5 Frequencies (top) and proportions (bottom) of Suspect IDs, Filler IDs, and No IDs in the one-person (showup), two-person, and six-person lineup conditions for target-present (TP) and target-absent (TA) lineups with repeated and unrepeated targets. The "–" symbols represent nonexistent filler data for showups

		TP counts			TA counts		
Repetition	Size	Suspect ID	Filler ID	No ID	Suspect ID	Filler ID	No ID
Repeated	1	245	–	5	11	–	239
	2	238	7	5	11	20	219
	6	206	38	6	4	57	189
Unrepeated	1	239	–	12	19	–	232
	2	218	25	8	11	46	194
	6	180	55	16	18	76	157
		TP proportions			TA proportions		
Repetition	Size	Suspect ID	Filler ID	No ID	Suspect ID	Filler ID	No ID
Repeated	1	0.98	–	0.02	0.04	–	0.96
	2	0.95	0.03	0.02	0.04	0.08	0.88
	6	0.82	0.15	0.02	0.02	0.23	0.76
Unrepeated	1	0.95	–	0.05	0.08	–	0.92
	2	0.87	0.10	0.03	0.04	0.18	0.77
	6	0.72	0.22	0.06	0.07	0.30	0.63

the 2-s study phase, as in Experiments 1 and 2. Second, the retrieval phase always contained high-similarity fillers. Note that unlike Experiment 2, in Experiment 3, the faces presented for study were never blurred.

Procedure

Before the experiment began, participants saw an instruction screen indicating that the faces they would study would repeat several times or not repeat at all. All other procedural variables were the same as in Experiments 1 and 2.

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 levels of repetitions are shown in Table 5. As in Experiment 2, a mirror effect is evident as memory accuracy declines from the repeated to the non-repeated conditions. That is, the hit rate (i.e., suspect ID rate in TP lineups) decreases and the false alarm rate (i.e., suspect ID rate in TA lineups) increases for all three lineup size conditions. At the same time, the No ID rate increases in TP lineups and the No ID rate decreases in all three lineup size conditions.

The ROC curves are shown in Fig. 7, and they indicate that discriminability decreased with increasing lineup size in both repetition conditions. Although the overall pattern demonstrates the showup superiority effect, consistent with the results of Experiment 1 and with prior findings, the same trend is not as apparent for the unrepeated condition. The ROC curves in Fig. 8, which use the estimated TA suspect ID rate, shows a clearer pattern. Now, both repetition conditions exhibit the showup superiority effect, and discriminability decreases as a function of lineup size (the deleterious lineup size effect).

We next estimated d'_{IG} in each condition by fitting a signal detection model, with details again presented in the Appendix (OSM). We began by constraining d'_{IG} to be equivalent for all six conditions (two repetition conditions by three lineup size conditions). Allowing it to differ as a function of repetition condition significantly improved the fit, $\chi^2(2) = 30.89$, $p < .001$. This result indicates that the mean d'_{IG} for the repetition condition (3.33) significantly exceeded that for the blurred condition (2.43). We next asked whether allowing d'_{IG} to vary within each repetition condition would significantly improve the fit. Indeed, doing so resulted in a significant improvement of the fit, $\chi^2(2) = 10.16$, $p < .001$ for the repeated condition and $\chi^2(2) = 12.66$, $p < .001$ for the non-repeated condition. For both conditions, discriminability was lowest in the lineup size 6 condition and highest in the lineup size 1 (showup) condition (Table 6).

General discussion

To ensure lineup fairness, researchers recommend that lineups be used in preference to showups, that at least five fillers should be included in the lineup, and that everyone in the lineup should be sufficiently similar that the suspect does not stand out (Wells et al., 2020). Yet a longstanding question concerns how to optimize lineup performance by choosing fillers appropriately. Recent research found that discriminability is optimized by using description-matched fillers who are

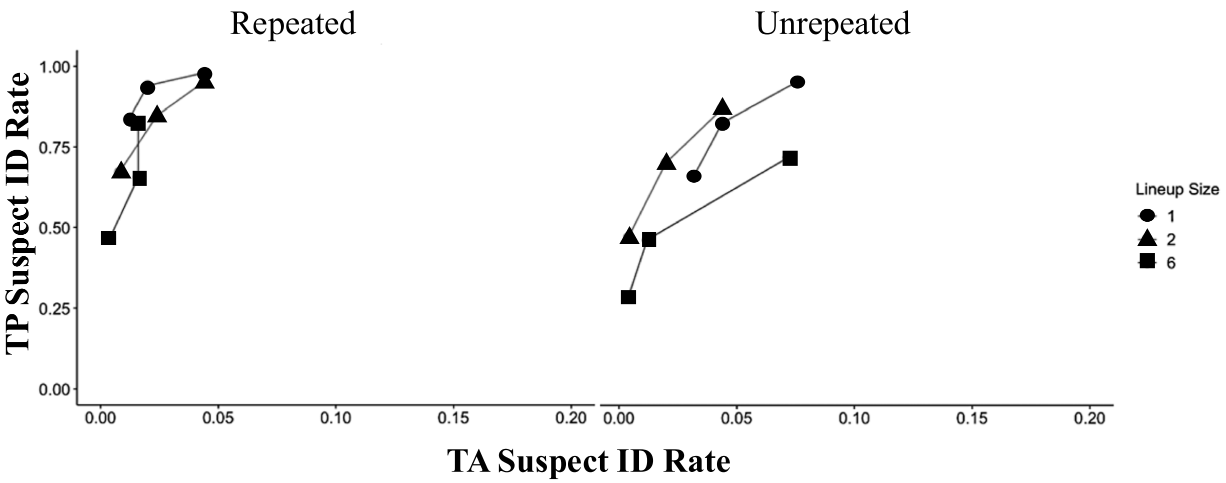

Fig. 7 Receiver operating characteristic (ROC) data from the repeated and unrepeated conditions of Experiment 3. The target-present (TP) suspect ID rate is the proportion of TP lineups that resulted in a suspect ID. The target-absent (TA) suspect ID rate is the proportion of designated innocent suspect chosen at lineup. Note that filler IDs from TP lineups are not represented in these ROC plots, but they are included when models are fit to the data

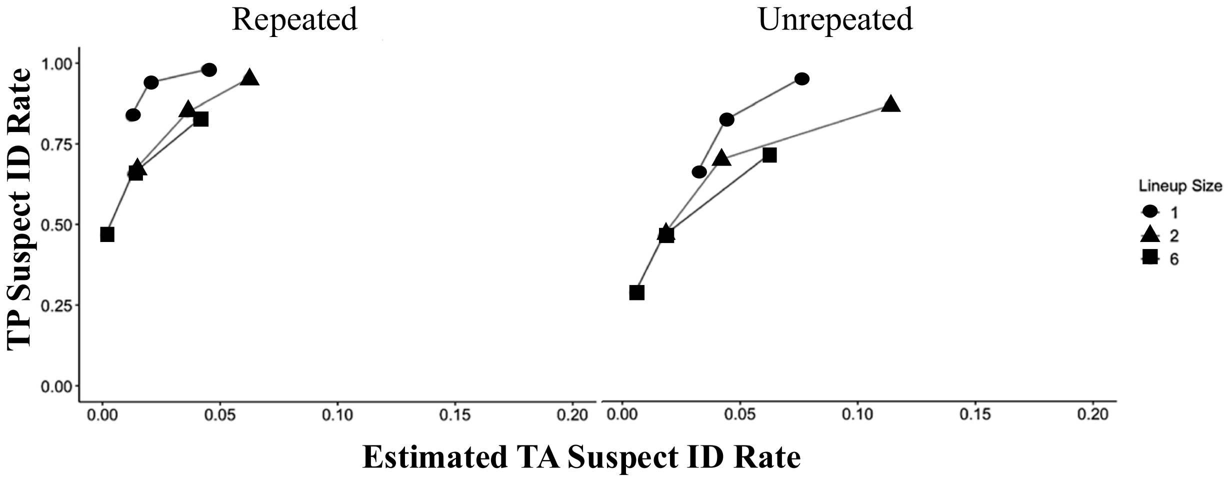

Fig. 8 Receiver operating characteristic (ROC) data from the repeated and unrepeated conditions of Experiment 3. The target-present (TP) suspect ID rate is the proportion of TP lineups that resulted in a suspect ID. The estimated target-absent (TA) suspect ID rate is the proportion of TA lineups that resulted in a filler ID divided by the lineup size of 1 or 2 or 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

otherwise maximally dissimilar to the suspect (Colloff et al., 2021; Shen et al., 2023, 2024), in agreement with an earlier study conducted long ago (Wells et al., 1993). Moreover, when low-similarity fillers are used, discriminability increases as lineup size increases. By contrast, when description-matched fillers are highly similar to the suspect, increasing lineup size *decreases* discriminability (Shen et al., 2023; Experiments 1 and 3 here). It is not yet clear why the effect of lineup size reverses when high-similarity fillers are used, but it is clear that the use of high-similarity fillers impairs overall discriminability (relative to when low-similarity fillers are used). Here, we investigated whether the effect of lineup would reverse as a function of overall discriminability when the relevant experimental manipulation occurs at encoding rather than at retrieval.

In Experiment 1, we replicated previously reported trends (Shen et al., 2024), reaffirming the main effect of filler similarity (a retrieval manipulation) on discriminability (Carlson et al., 2019; Colloff et al., 2021). Experiments 2 and 3 then varied overall discriminability across conditions using an encoding manipulation instead. In Experiment 2, we found that when low-similarity fillers are used, increasing lineup size increased discriminability whether encoding conditions were favorable or unfavorable. In Experiment 3, we found that when high-similarity fillers were used, increasing lineup size decreased discriminability whether encoding conditions were favorable or unfavorable. Indeed, in this experiment, the two lineup conditions yielded lower discriminability than the showup condition. This constitutes an exception to the longstanding assumption that lineups are always better than showups.

Table 6 Discriminability estimates from Experiment 3, with d'_{IG} free to vary as a function of lineup size

Repetition	Size	d'_{IG}
Repeated	1	3.61
	2	3.24
	6	3.19
Unrepeated	1	2.81
	2	2.58
	6	2.56

How can we make sense of these findings theoretically? As noted earlier, a model known as the Ensemble model correctly predicts that d'_{IG} will increase as filler similarity decreases (an effect that is reliably observed) and also predicts that d'_{IG} will increase as lineup size increases (an effect that is observed only when low-similarity fillers are used). The Ensemble model makes that lineup-size prediction because it assumes that identification decisions are based on how familiar a face is relative to the average familiarity of the faces in the lineup. The average level of familiarity in a lineup is more precisely determined as the number of faces in the lineup increases (thereby reducing noise). From the perspective of this model, the issue to be explained is why the effect of lineup size reverses when high-similarity fillers are used.

Shen et al. (2024) proposed that a noise factor may come into play as the task becomes more difficult. As an example, repeated back-and-forth eye movements might increase as the task becomes more difficult with increasing filler similarity. Any such effect might become increasingly pronounced as lineup size increases (e.g., Utochkin et al., 2023). However, the results of Experiments 2 and 3 here suggest that a comparable effect is not observed when the lineup task becomes harder due to manipulations at encoding. In other words, even when overall performance was poor because the faces at encoding were blurry, when low-similarity fillers were used, discriminability still increased with increasing lineup size. By contrast, even when overall performance was high because the faces at encoding were repeated, when high-similarity fillers were used, discriminability still decreased with increasing lineup size. It thus appears that the eye-movement account may not be correct, although a more definitive test would be provided by a future investigation that uses eye-tracking to directly measure eye movements.

Given that the variable effects of increasing lineup size appear to be determined by the conditions at retrieval, it seems reasonable to ask what it might be about highly similar fillers that reduces discriminability as lineup size increases. Although this effect is opposite in direction to the effect predicted by the Ensemble model, it may nevertheless also be related to ensemble perception. According to the Ensemble model, the representation of a similar set of faces creates an ensemble (average) representation that essentially reflects the shared features of the faces in the lineup. These shared features are nondiagnostic of the perpetrator, precisely because they are shared (Wixted & Mickes, 2014). According to this model, each face is assessed in terms of the degree to which its familiarity differs from the average familiarity associated with the ensemble representation (subtracting away the portion of the memory signal generated by shared features), which enhances the role played by the diagnostic (i.e., non-shared) features. In essence, any noise contributed by the shared features is subtracted away.

At least when low-similarity fillers are used, the enhancement of discriminability with increasing lineup size theoretically occurs because the average signal that is subtracted away (i.e., the ensemble) is more precisely determined. However, as the faces in the lineup become increasingly similar to each other, increasing the number of shared features while simultaneously decreasing the number of potentially diagnostic features, the ensemble representation becomes ever more complete. As lineup size increases, it becomes ever more precise as well. When that occurs, it may serve to command attention to the shared features. Indeed, some research suggests that the ensemble is perceived even before the individual items are represented (Haberman et al., 2009; Im et al., 2021; Liu et al., 2023). If so, it is possible that the ensemble representation, in addition to providing a convenient means of subtracting away the effect of shared features, also biases perception toward the shared facial features in the lineup.

Any such attention-capturing effect would reduce the role played by the diagnostic features because attention has been drawn away from them, reducing discriminability. Now, when filler similarity is high, increasing lineup size would be problematic because the more complete ensemble representation would also be more precisely specified. Under such conditions, all of the faces in the lineup would look more alike than they would in the absence of an ensemble-induced attention-guiding effect. This bottom-up biasing would interfere with the top-down goal of subtracting away the ensemble to focus on diagnostic facial features in order to complete the task-relevant goal of identifying the target face in the lineup.

The proposed mechanistic explanation seems plausible to us assuming the quality of the ensemble representation depends on set size and similarity. The ensemble perception literature has sometimes reported an increase in ensemble sensitivity with set size (Alvarez, 2011; Ariely 2001; Chong et al., 2008; Kanaya et al., 2018; Robitaille & Harris, 2011; Utochkin et al., 2023), and an increase in ensemble precision with less variance (i.e., higher similarity) (Corbett et al., 2012; Goldenberg et al., 2020; Maule & Franklin, 2015; Peng, 2021). It seems reasonable to us to speculate that the ensemble representation may increasingly guide attention to faces the more precise and complete it is. And the more it draws attention to shared (nondiagnostic) features due to this attentional capture (Theeuwes, 1992, in press), the greater the interference with the top-down objective of identifying diagnostic features.

Methodologically, we manipulated filler similarity using a morphing procedure, but it seems likely that artificial intelligence (AI) image generation algorithms will make this job much easier in the near future. Indeed, AI is already making its way into eyewitness identification research (e.g., Bell et al., 2024; Kleider-Offutt et al., 2024), a trend that one can only imagine will accelerate. Research on the effects of filler similarity on eyewitness identification may accelerate accordingly.

Some limitations of the experiments reported here may already be apparent but are worth highlighting, nonetheless. First, the work reported here was designed to enhance our understanding of lineup decisions in terms of underlying memory signals, not to directly inform policymakers about what kind of lineup formats yield the best applied outcomes. For example, we used morphed stimuli, which are not used in real police lineups, and we tested participants multiple times, which is typical of experiments in cognitive psychology but differs from the real world, where witnesses are usually tested only once. In addition, thus far, the apparent interaction between filler similarity and lineup size on d'_{IG} has only been observed in our lab. Going forward, it will be important for the interaction to be independently replicated before treating it as established knowledge (Wilson et al., 2020, 2022). And finally, for reasons unknown, the innocent suspect was less attractive than the fillers in the two-person lineups throughout our three experiments (see Tables 1, 3, and 5). Note that this apparent stimulus artifact was true for both the high- and low-similarity conditions and thus does not account for the opposite patterns observed in those two conditions with respect to lineup size. Still, it was an unexpected effect and is therefore another limitation of this research.

The new theoretical perspective advanced here was conceived in light of an unexpected pattern of results and is therefore tentative. One way to test it might be to instruct participants in the high-similarity condition as to what the diagnostic and non-diagnostic features are. Armed with that knowledge, it seems reasonable to hypothesize that they could better resist the attractive attentional force of the ensemble representation and better maintain attention to the more useful diagnostic features. In that case, perhaps increasing lineup size would enhance discriminability (instead of harming it) even when high-similarity fillers are used.

Lineup research has been conducted at an accelerated pace since at least the 1970s. However, in light of the importance of police lineups in the real world, gaining a better theoretical understanding of how to use fillers to optimize the performance of eyewitnesses still seems overdue. The work reported here constitutes one contribution to the theoretical importance of this line of research.

Supplementary information The online version contains supplementary material available at https://doi.org/10.3758/s13421-024-01649-x.

Funding Not applicable.

Data availability Materials/data are available via the Open Science Framework at: https://osf.io/5z4y7/.

Code availability Not applicable.

Declarations

Ethics approval This research was approved by the university's internal review board (protocol # 121186).

Consent to participate All participants provided informed consent.

Consent for publication All authors provided consent for publication.

Open practices statement The data and materials for all experiments are available via the Open Science Framework at: https://osf.io/5z4y7/ and Experiment 1 was preregistered.

Conflicts of interest The authors have no conflicts of interest.

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. https://doi.org/10.1037/xap0000340

Alvarez, G. A. (2011). Representing multiple objects as an ensemble enhances visual cognition. Trends in cognitive sciences, 15(3), 122–131. https://doi.org/10.1016/j.tics.2011.01.003

Ariely, D. (2001). Seeing sets: Representation by statistical properties. *Psychological Science, 12*, 157–162. https://doi.org/10.1111/1467-9280.00327

Bell, R., Menne, N. M., Mayer, C., & Buchner, A. (2024). On the advantages of using AI-generated images of filler faces for creating fair lineups. *Scientific Reports, 14*(1), 12304. https://doi.org/10.1038/s41598-024-63004-z

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), 1–16. https://doi.org/10.1186/s41235-019-0172-5

Chong, S. C., Joo, S. J., Emmmanouil, T. A., & Treisman, A. (2008). Statistical processing: Not so implausible after all. *Perception & Psychophysics, 70*(7), 1327–1334. https://doi.org/10.3758/pp.70.7.1327

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

Colloff, M. F., & Wixted, J. T. (2020). Why are lineups better than showups? A test of the filler siphoning and enhanced discriminability accounts. *Journal of Experimental Psychology: Applied, 26*(1), 124. https://doi.org/10.1037/xap0000218

Corbett, J. E., Wurnitsch, N., Schwartz, A., & Whitney, D. (2012). An aftereffect of adaptation to mean size. *Visual Cognition, 20*(2), 211–231. https://doi.org/10.1080/13506285.2012.657261

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. https://doi.org/10.1037/lhb0000095

Glanzer, M., & Adams, J. K. (1990). The mirror effect in recognition memory: Data and theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*, 5–16. https://doi.org/10.1037/0278-7393.16.1.5

Glanzer, M., Adams, J. K., Iverson, G. J., & Kim, K. (1993). The regularities of recognition memory. *Psychological Review, 100*, 546–567. https://doi.org/10.3758/s13421-018-0864-y

Goldenberg, A., Sweeny, T. D., Shpigel, E., & Gross, J. J. (2020). Is this my group or not? The role of ensemble coding of emotional expressions in group categorization. *Journal of Experimental Psychology: General, 149*(3), 445. https://doi.org/10.1037/xge0000651

Haberman, J., Harp, T., & Whitney, D. (2009). Averaging facial expression over time. *Journal of Vision, 9*(11), 1–13. https://doi.org/10.1167/9.11.1

Im, H. Y., Cushing, C. A., Ward, N., & Kveraga, K. (2021). Differential neurodynamics and connectivity in the dorsal and ventral visual pathways during perception of emotional crowds and individuals: A MEG study. *Cognitive, Affective & Behavioral Neuroscience, 21*(4), 776–792. https://doi.org/10.3758/s13415-021-00880-2

Kanaya, S., Hayashi, M. J., & Whitney, D. (2018). Exaggerated groups: Amplification in ensemble coding of temporal and spatial features. *Proceedings of the Royal Society b: Biological Sciences, 285*(1879), 20172770. https://doi.org/10.1098/rspb.2017.2770

Kleider-Offutt, H., Stevens, B., Mickes, L., & Boogert, S. (2024). Application of artificial intelligence to eyewitness identification. *Cognitive Research: Principles and Implications, 9*(1), Article 19. https://doi.org/10.1186/s41235-024-00542-0

Liu, R., Ye, Q., Hao, S., Li, Y., Shen, L., & He, W. (2023). The relationship between ensemble coding and individual representation of crowd facial emotion. *Biological Psychology, 180*, 108593. https://doi.org/10.1016/j.biopsycho.2023.108593

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. https://doi.org/10.1007/BF01044829

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. https://doi.org/10.3758/s13428-014-0532-5

Mazyar, H., Van den Berg, R., & Ma, W. J. (2012). Does precision decrease with set size? *Journal of Vision, 12*(6), 10–10. https://doi.org/10.3758/s13428-014-0532-5

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. https://doi.org/10.1167/13.5.8

McClelland, J. L., & Chappell, M. (1998). Familiarity breeds differentiation: A subjective-likelihood approach to the effects of experience in recognition memory. *Psychological Review, 105*, 724–760. https://doi.org/10.1037/0033-295X.105.4.734-760

Maule, J., & Franklin, A. (2015). Effects of ensemble complexity and perceptual similarity on rapid averaging of hue. *Journal of Vision, 15*(4), 6–6. https://doi.org/10.1167/15.4.6

Mickes, L., Seale-Carlisle, T. M., Chen, X., & Boogert, S. (2023). pyWitness 1.0: A Python eyewitness identification analysis toolkit. *Behavior Research Methods*, https://doi.org/10.3758/s13428-023-02108-2

National Institute of Justice. (1999). *Eyewitness evidence: A guide for law enforcement*. Washington, DC: Office of Justice Programs, U.S. Department of Justice.

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. https://doi.org/10.1037/lhb0000272

Osth, A. F., Dennis, S., & Heathcote, A. (2017). Likelihood ratio sequential sampling models of recognition memory. *Cognitive Psychology, 92*, 101–126. https://doi.org/10.1037/lhb0000272

Peng, S., Liu, C. H., & Hu, P. (2021). Effects of subjective similarity and culture on ensemble perception of faces. *Attention, Perception, & Psychophysics, 83*, 1070–1079. https://doi.org/10.3758/s13414-020-02133-9

Robitaille, N., & Harris, I. M. (2011). When more is less: Extraction of summary statistics benefits from larger sets. *Journal of Vision, 11*(12), 18–18. https://doi.org/10.1167/11.12.18

Semmler, C., Dunn, J., Mickes, L., & Wixted, J. T. (2018). The role of estimator variables in eyewitness identification. *Journal of Experimental Psychology: Applied, 24*(3), 400–415. https://doi.org/10.1037/xap0000157

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. https://doi.org/10.1037/rev0000408

Shen, K. J., Huang, J., Lam, A., & Wixted, J. T. (2024). The effects of filler similarity and lineup size on eyewitness identification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. https://doi.org/10.1037/xlm0001342

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM: Retrieving Effectively from Memory. *Psychonomic Bulletin & Review, 4*, 145–166. https://doi.org/10.3758/bf03209391

Stretch, V., & Wixted, J. T. (1998). Decision rules for recognition memory confidence judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*, 1397–1410. https://doi.org/10.1037/0278-7393.24.6.1397

Theeuwes, J. (1992). Perceptual selectivity for color and form. *Perception & Psychophysics, 51*(6), 599–606. https://doi.org/10.3758/BF03211656

Theeuwes, J. (in press). Attentional capture and control. *Annual Review of Psychology*.

Utochkin, I. S., Choi, J., & Chong, S. C. (2023). A population response model of ensemble perception. *Psychological Review, 131*(1), 36–57. https://doi.org/10.1037/rev0000426

Wells, G. L. (1978). Applied eyewitness-testimony research: System variables and estimator variables. *Journal of Personality and Social Psychology, 36*(12), 1546. https://doi.org/10.1037/0022-3514.36.12.1546

Wells, G. L., Rydell, S. M., & Seelau, E. P. (1993). The selection of distractors for eyewitness lineups. *Journal of Applied Psychology, 78*(5), 835. https://doi.org/10.1037/0021-9010.78.5.835

Wells, G. L., Small, M., Penrod, S., Malpass, R. S., Fulero, S. M., & Brimacombe, C. E. (1998). Eyewitness identification procedures: Recommendations for lineups and photospreads. *Law and Human Behavior, 22*(6), 603. https://doi.org/10.1023/A:1025750605807

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*(1), 3. https://doi.org/10.1037/lhb0000359

Wilson, B. M., Harris, C. R., & Wixted, J. T. (2020). Science is not a signal detection problem. *Proceedings of the National Academy of Sciences of the United States of America, 117*(11), 5559–5567. https://doi.org/10.1073/pnas.1914237117

Wilson, B. M., Harris, C. R., & Wixted, J. T. (2022). Theoretical false positive psychology. *Psychonomic Bulletin & Review, 29*(5), 1751–1775. https://doi.org/10.3758/s13423-022-02098-w

Wixted, J. T., & Gaitan, S. (2002). Cognitive theories as reinforcement history surrogates: The case of likelihood ratio models of human recognition memory. *Animal Learning & Behavior, 30*, 289–305. https://doi.org/10.3758/BF03195955

Wixted, J. T., & Mickes, L. (2014). A signal-detection-based diagnostic-feature-detection model of eyewitness identification. *Psychological Review, 121*(2), 262. https://doi.org/10.1037/a0035940

Wixted, J. T., Vul, E., Mickes, L., & Wilson, B. M. (2018). Models of lineup memory. *Cognitive Psychology, 105*, 81–114. https://doi.org/10.1016/j.cogpsych.2018.06.001

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. https://doi.org/10.1002/acp.3644

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.