

# Journal of Experimental Psychology: Applied

## The Effect of Lineup Size on Eyewitness Identification

Melisa Akan, Maria M. Robinson, Laura Mickes, John T. Wixted, and Aaron S. Benjamin

Online First Publication, December 3, 2020. <http://dx.doi.org/10.1037/xap0000340>

### CITATION

Akan, M., Robinson, M. M., Mickes, L., Wixted, J. T., & Benjamin, A. S. (2020, December 3). The Effect of Lineup Size on Eyewitness Identification. *Journal of Experimental Psychology: Applied*. Advance online publication. <http://dx.doi.org/10.1037/xap0000340>

# The Effect of Lineup Size on Eyewitness Identification

Melisa Akan<sup>1</sup>, Maria M. Robinson<sup>1</sup>, Laura Mickes<sup>2</sup>, John T. Wixted<sup>3</sup>, and Aaron S. Benjamin<sup>1</sup>

<sup>1</sup> Department of Psychology, University of Illinois at Urbana–Champaign

<sup>2</sup> School of Psychological Science, University of Bristol

<sup>3</sup> Department of Psychology, University of California, San Diego

Eyewitness identification via lineup procedures is an important and widely used source of evidence in criminal cases. However, the scientific literature provides inconsistent guidance on a very basic feature of lineup procedure: lineup size. In two experiments, we examined whether the number of fillers affects diagnostic accuracy in a lineup, as assessed with *receiver-operating characteristic* (ROC) analysis. Showups (identification procedures with one face) led to lower discriminability than simultaneous lineups. However, in neither experiment did the number of fillers in a lineup affect discriminability. We also evaluated competing models of decision-making from lineups. This analysis indicated that the standard *Independent Observations* (IO) model, which assumes a decision rule based on the comparison of memory strength signals generated by each face in a lineup, is incapable of reproducing the lower level of performance evident in showups. We could not adjudicate between the *Ensemble* model, which assumes a decision rule based on the comparison of the strength of each face with the mean strength across the lineup, and a newly introduced *Dependent Observations* model, which adopts the same decision rule as the IO model, but with correlated signals across faces. We draw lessons for users of lineup procedures and for basic research on eyewitness decision making.

## Public Significance Statement

The current set of studies suggest that showups (an identification procedure with a single face) lead to a reduced ability to discriminate a guilty suspect from an innocent suspect compared with lineups—simultaneous presentation of the suspect with other faces known to be innocent (fillers). The number of faces presented along with the suspect in a lineup does not have a measurable impact on this ability. These results can inform practice in law enforcement: There may be little gain in seeking to include fillers in a lineup beyond those that are well matched and immediately available, especially if the process of finding those fillers will delay the eyewitness identification procedure considerably.


**Keywords:** eyewitness memory, lineup size, ROC analysis, signal detection-based models, showups

Eyewitness identifications constitute a major source of evidence used by the criminal justice system in the process of determining

the perpetrator of a crime. However, memory is faulty and identification decisions are suggestible and prone to bias. As revealed by DNA analysis, eyewitness identifications of innocent suspects have played a role in the majority of wrongful convictions (Innocence Project, 2018). In the recent decades, there has been an increasing effort by social scientists to help improve the methods and procedures used by the law enforcement to collect eyewitness evidence.

Ongoing research tries to uncover the best practices surrounding the use of photo lineups—one of the most commonly used procedures in the United States (Police Executive Research Forum, 2013). In a photo lineup, a suspect—who may or may not be the actual perpetrator—is presented together with a number of fillers, who are known to be not guilty of the investigated crime. In this procedure, eyewitnesses are tasked with identifying the perpetrator if the perpetrator happens to be in the lineup, and if not, to reject the lineup. If eyewitnesses had perfect memory, they would always correctly identify the guilty suspect if they were in the lineup and they would always correctly reject a lineup with an innocent suspect. Put differently, having good memory of the perpetrator should lead to a greater ability to discriminate between guilty and

Melisa Akan  <https://orcid.org/0000-0001-8811-6144>

Maria M. Robinson  <https://orcid.org/0000-0002-9797-4068>

Maria M. Robinson is now at the Department of Psychology, University of California, San Diego.

Part of this work was supported by the National Science Foundation Grant SES-1155248 to John T. Wixted and Laura Mickes. The content is solely the responsibility of the authors and does not necessarily reflect the views of the National Science Foundation. We thank Alex Wooten for sharing the data set reported in Wooten et al. (2000), Jessica Siler for her help with editing the stimuli, and the members of the Human Memory and Cognition Lab for their feedback throughout the execution of this work.

The raw data files for both experiments as well as the code used to construct and test the models are available on the OSF page of this study and can be accessed at <https://osf.io/xcfhj/>.

Correspondence concerning this article should be addressed to Melisa Akan, Department of Psychology, University of Illinois at Urbana–Champaign, 603 East Daniel Street, Champaign, IL 61820, United States. Email: [makan2@illinois.edu](mailto:makan2@illinois.edu)

innocent suspects. A major goal of research on eyewitness memory is to find lineup procedures that are associated with greater discriminability between innocent and guilty suspects (Clark et al., 2015). This goal requires the application of a unified theory of discriminability and decision making, like *signal detection theory* (SDT). The application of SDT to the problems of eyewitness memory has had much success in recent years (e.g., Gronlund & Benjamin, 2018; Wixted & Mickes, 2012, 2015).

Many aspects of lineup procedures have been studied, including the composition of a lineup (e.g., the characteristics of fillers, e.g., Wells et al., 1993), the presentation method of the photo array (simultaneous vs. sequential, e.g., Mickes et al., 2012), and the position of the suspect in the lineup (e.g., Carlson et al., 2008). There is relatively less research dedicated to the size of the lineup. Different countries have different standards for the number of fillers included in a typical lineup. In the United States, police departments mostly use five fillers (six-person lineups; Police Executive Research Forum, 2013), the United Kingdom typically uses eight fillers (Police & Criminal Evidence Act, 1984), and in Canada, the recommended lineup consists of at least nine fillers (Report on the Prevention of Miscarriages of Justice, 2005). In this study, we investigate whether and how lineup size affects discriminability. We will return shortly to what we mean by discriminability; it is important to note that this is a different criterion than that used in most prior work on lineup size (and by many researchers in eyewitness memory more generally).

The historical analytic approach taken by the few studies examining lineup size treats correct identifications and false identifications separately. The correct identification, or *hit*, rate (HR) refers to the endorsement rate for the suspect in lineups in which the suspect is guilty (target present [TP] lineups). The false identification, or *false alarm*, rate (FAR), refers to the endorsement rate for the suspect in lineups in which the suspect is innocent (target absent [TA] lineups). Generally, in lab studies that designate an innocent suspect, the fillers are held constant across TA and TP lineups, and the target is replaced by another specific filler (i.e., the innocent suspect) in TA lineups. Note that identifications of lineup members who are known to be innocent (i.e., fillers) are not considered as false alarms within this framework. In studies that do not designate a single innocent suspect, the FAR is estimated by dividing the total false identification rate in TA lineups by lineup size (e.g., Mickes, 2015).

### Measurement of Accuracy in Lineup Identifications

We can know for certain that a particular lineup procedure leads to greater discriminability if it simultaneously leads to an increase in hit rate and a decrease in false alarm rate. This is a phenomenon that is common enough in research on recognition memory to have earned its own name—the *mirror effect* (Glanzer & Adams, 1985). However, it is often the case that the comparison of different lineup procedures (e.g., simultaneous vs. sequential presentation, Steblay et al., 2001) leads to these two values increasing or decreasing in tandem across conditions. That is, a lineup procedure that leads to an increase in correct identifications of the guilty suspect is also likely to lead to an increase in false identifications of the innocent suspect (for a review, see Clark, 2012<sup>1</sup>). We will return shortly to an explanation of why this occurs; right now, we

confront the more pressing problem: In this case, how do we know which procedure is associated with higher discriminability?

One attempt to capture changes in HRs and FARs in a single measure is reflected in the *diagnosticity ratio*: the ratio of correct identifications to false identifications ( $HR \div FAR$ ; e.g., Wells & Lindsay, 1980). Some emphasize the practical significance of this measure, as it describes the odds of an identification being correct given that a suspect identification was made (e.g., Wells & Olson, 2002). Additionally, on its face, this measure has appealing qualities: When hits go up, the diagnosticity ratio goes up, and as false alarms go up, the diagnosticity ratio goes down. However, combining HR and FAR in this manner embodies the strong theoretical position that all (HR, FAR) pairs that lead to equivalent ratios—say (1.0, 0.5), (0.8, 0.4), and (0.2, 0.1)—reflect equivalent latent discriminability between guilty and innocent suspects (Rotello et al., 2015). This claim has not been supported by analyses of *receiver-operating characteristics* (ROC), which plot (HR, FAR) pairs across conditions that are *known* to lead to equivalent discriminability (Swets, 1986b).

The failure of the diagnosticity ratio is not unexpected; it is known from decades of research in perception and memory that the ROC has a prototypical shape from which it rarely deviates, and that that shape is inconsistent with the one predicted by diagnosticity ratio as an index of discriminability (Swets, 1986a). Of particular concern is how the actual shape of the ROC biases the diagnosticity ratio. As responding becomes more conservative, the diagnosticity ratio increases (refer to Table 1 in Wixted & Mickes, 2012). That is, even when latent discriminability is constant, differing response biases will lead to different values of the diagnosticity ratio (Wixted & Mickes, 2014). Consequently, a lineup procedure that leads to more conservative responding (lower HR and FAR values) will be misinterpreted as having greater diagnostic accuracy.

Almost all studies that manipulated lineup size to this point have used only a single HR-FAR pair associated with each lineup procedure. Consequently, they used one or both of two analytically imperfect procedures to make inferences about the diagnostic accuracy of different lineup procedures: (a) analyzing HRs and FARs separately, or (b) using an inappropriate transformation of HR and FAR values, like the diagnosticity ratio. Among the few studies that have examined lineup size, some did not detect a statistical difference in either correct identification rate from TP lineups or the filler identification rate (identification of any lineup member) from TA lineups (e.g., Brewer et al., 2006; Nosworthy & Lindsay, 1990; Pozzulo et al., 2010).<sup>2</sup> This pattern has sometimes been cited in support of the inference that larger lineup sizes are more reliable than the smaller ones, because larger lineups appear

<sup>1</sup> The simultaneous increase in correct and false identification rates was observed for all comparisons except for the comparison of showups with lineups. The lineup procedure was associated with both lower false ID rates and slightly higher correct ID rates compared WITH showups, averaging across 15 comparisons reported in the eyewitness memory literature.

<sup>2</sup> None of these studies were amply powered; the sample size for each lineup condition ranged between 22 to 30, with each participant making only a single lineup identification.

to provide greater protection for innocent suspects while not causing a detectable detriment on the identification of the guilty suspect. Even in cases where hit rates decrease with lineup size, that decrease was seen to be mitigated by the larger decrease in false alarm rates. The underlying idea is that increasing lineup size will be beneficial as long as the proportion increase in size exceeds the resulting increase in choosing rates—the probability of any lineup member being identified—from TA lineups (e.g., Levi & Lindsay, 2001).

Following this logic, Levi (2007, 2012) examined identification performance using exceptionally large lineups. In both studies, there was a reduction in FARs accompanied by either no detectable change in HR (Levi, 2007) or a reduction in HR (Levi, 2012) with increasing lineup size. For example, Levi (2012) compared a lineup of size 12 with a lineup of size 120 and found a difference in both hit rates (23% vs. 10%, respectively) and in the combination of false alarm and filler identification rates in TA lineups (34% vs. 53%, respectively). Even though an innocent suspect was designated, the FARs were estimated by dividing the choosing rate by lineup size in TA lineups ( $34\% \div 12 = 2.8\%$  for lineup size of 12, and  $53\% \div 120 = 0.4\%$  for lineup size of 120). Based on these hit and false alarm rates, Levi computed the probability that the suspect was innocent, conditional upon having been endorsed for each lineup procedure (i.e.,  $\text{FAR} \div [\text{HR} + \text{FAR}]$ ). This probability was  $(2.8 \div [23 + 2.8]) = 10.9\%$  for the 12-person lineup and  $(0.4 \div [10 + 0.4]) = 3.8\%$  for the 120-person lineup, leading to the conclusion that the latter procedure should be preferred over the former.

This approach embodies several errors: (a) confounding of discriminability with bias by basing inferences on a single HR-FAR pair, (b) acceptance of the null hypothesis by drawing the inference that increasing lineup size does *not* affect correct identification rates, and (c) disregarding correct identification rates (i.e., convicting the guilty suspect) while exclusively focusing on reducing false identifications (i.e., exonerating the innocent suspect). Importantly, across studies, lineups that contain 24 or more members appear to lead to lower correct identification rates (10–18%; Levi, 2007, 2012) than smaller lineups (21–81%; Brewer et al., 2006; Nosworthy & Lindsay, 1990; Pozzulo et al., 2010). It is difficult to know exactly what to make of this literature when all of these concerns are taken into account.

Clearly, a problem with these analytic approaches is the lack of consideration given to the confounding effects of response bias. Some experimental conditions lead to a greater willingness to endorse, and it is this tendency that needs to be extricated from our measures before we can assess the true diagnosticity of a procedure. This can be accomplished by simultaneously considering multiple (HR, FAR) pairs across varying levels of response bias associated with a single lineup procedure (Clark et al., 2015; Wixted & Mickes, 2012). A convenient method of obtaining multiple (HR, FAR) pairs is the use of confidence ratings as a proxy for criteria placed at multiple levels of response bias. The (HR, FAR) point associated with the most conservative response bias is based on identification responses made with the highest level of confidence, and the most liberal point includes all identification responses encompassing any level of confidence (Macmillan & Creelman, 2005; Wickens, 2002). An ROC constructed with this method allows us to decide which procedure is diagnostically superior to the other. This approach is widely used in

psychology (e.g., Egan, 1958; Ratcliff et al., 1992), in diagnostic medicine (e.g., Lusted, 1971; Beck, 1991), and in eyewitness memory (e.g., Mickes et al., 2012; Wixted, 2020). Figure 1 shows an example of how these confidence categories can be used to produce an ROC function.

Three empirical studies have used ROC analysis to evaluate how showups—presentation of the suspect alone, without any fillers—compare with lineups. Showups are typically used if a suspect is apprehended within a short time interval after the crime, and within a close distance from the crime scene (National Research Council, 2015). Otherwise, the showup procedure is considered to be “inherently suggestive” (*Stovall v. Denno*, 1967) and does not provide sufficient protection for an innocent suspect. All three studies demonstrated that showups are associated with lower discriminability (Gronlund et al., 2012; Mickes, 2015; Wetmore et al., 2017). These results suggest that the size of the lineup may be an important variable, meriting this same kind of discrimination analysis. A recent study by Wooten et al. (2020) addressed this problem by examining ROCs across showups and lineups of varying sizes with a large sample size. Consistent with prior work, showups led to worse discriminability compared with lineups; however, increasing lineup size did not lead to significant increases in discriminability. In the present study, we examined the effect lineup size on discriminability in two experiments, and evaluated two signal-detection theoretic models in terms of how well they characterize empirical data from these experiments. We also used data from Wooten et al. (2020) to conduct confirmatory model recovery and cross-validation analyses. The larger sample size used by Wooten et al. (2020) made this data set a better candidate for conducting model recovery simulations because the recovered parameter estimates are likely to be more reliable.

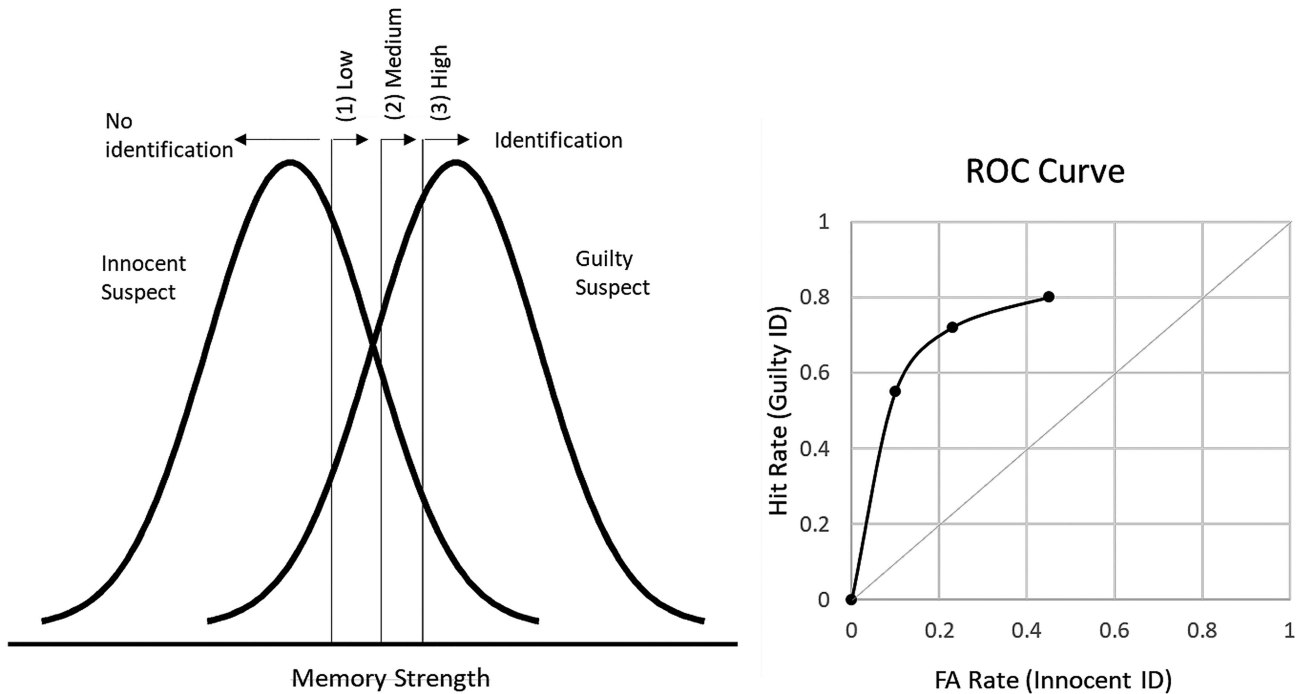
The type of models we bring to the evaluation of lineup performance here are descendants of tried-and-true models in research in memory and perception. Studies from the recognition memory literature have shown that forced-choice recognition performance decreases with increasing number of distractors (Kintsch, 1968; Postman, 1950), a finding that follows from predictions of a model in which stimuli are evaluated independently and the stimulus with the maximum memory strength value is chosen. This model, which we introduce shortly as the *Independent Observations* (IO) MAX model, can be applied to lineups with only minor modifications. We also consider an alternative detection-theoretic model, the *Ensemble* model, in which each face is compared to the mean strength across all of the faces in the lineup (Wixted et al., 2018). Before presenting the current experiments and the models, we first evaluate the use of ROC analysis in examining discriminability associated with varying lineup sizes.

### Receiver-Operating Characteristic (ROC) Analyses

As mentioned earlier, the comparison of overall single hit and false alarm pairs is not suitable for shedding light on the question of discriminability because a single (HR, FAR) value pair is determined by both underlying discriminability and response bias. In the particular case of different lineup sizes, there are two additional complications.

**Figure 1**

*Latent Probability Distributions of Memory Strength for the Innocent and Guilty Suspects and the Corresponding ROC Curve*



*Note.* Left panel: memory strength probability distributions yielded by a guilty and an innocent suspect. The greater the memory strength, the more evidence of guilt. If eyewitnesses are asked to rate their level of confidence on a 3-point scale, the vertical lines represent the level of memory strength required to make an identification response with low, medium, and high confidence. The hit and false alarm rates associated with each level of confidence correspond to the area under the distributions for the guilty suspect and the innocent suspect, respectively, to the right of the vertical line. The right panel depicts an receiver-operating characteristic (ROC) curve that corresponds to the (FAR, HR) values associated with each of the three levels of confidence. The diagonal represents chance performance. FA = false alarm.

The lineup size manipulation will inevitably affect hit and false alarm rates, even if it has no direct effect on discriminability or response bias. This is because, even assuming equal underlying discriminability, the likelihood of an identification response landing on a guilty or an innocent suspect will decrease with an increase in the number of lineup members, a phenomenon that has been referred to as “shifting” (Wells, 1993), and, more recently, as “filler siphoning” (Smith et al., 2017; Wells et al., 2015). If a truly innocent suspect is no more familiar than the fillers in a TA lineup, the likelihood of the innocent suspect being identified as guilty will be reduced as lineup size increases. This occurs because, as more fillers are added to a lineup, the probability that one of those fillers will serendipitously have higher familiarity than the innocent suspect will increase. Notably, it is also the case that fillers can “siphon” responses away from the target in a TP trial. In a TP lineup, the likelihood of one of the fillers eliciting greater familiarity than the guilty suspect will also increase as more fillers are added to a lineup. Thus, we would also expect a reduction in hit rates as the number of fillers increase. The effect of lineup size on hit and false alarm rates even when the underlying discriminability and response bias are held constant reveals why it is so important to have a diagnostic tool, like ROC analysis, that takes both HR and FAR into account simultaneously.

The second problem concerns the range of possible FAR values, which is narrower with larger lineup sizes. In a fair lineup, no lineup member stands out from the rest and therefore, the maximum possible FAR is  $\sim 1/n$ , where  $n$  is the lineup size. Given the effect of lineup size on hit and false alarm rates, as well as the narrower range of the FAR values even when discriminability is held constant, one might question the utility of the ROC analysis (Lampinen, 2016).

This concern has been addressed by prior work with simulations of identification decisions based on lineups of different sizes, holding discriminability constant across procedures (Rotello & Chen, 2016; Wixted & Mickes, 2015). The simulations were based on a simple decision strategy known as the MAX decision rule in perception research (e.g., Eckstein et al., 2000) and the *best-above-criterion* decision rule in eyewitness memory research (Clark et al., 2011). According to this rule, an eyewitness identifies the face from a lineup that elicits the highest level of familiarity (i.e., the greatest match to the memory of the perpetrator) *if and only if* that level of familiarity exceeds the eyewitness’ internal response criterion for endorsement. This decision rule is based on a simple signal detection model in which memory signals from lineup members are assessed independently of one another. Hence, this model is named the *Independent Observations* model (Duncan, 2006; Macmillan & Creelman, 1991). Using this simple decision



model, Wixted and Mickes (2015) simulated data for the showup and six-person lineup procedures, and Rotello and Chen (2016) simulated data for the showup, two-, three-, and six-person lineups. Both simulations produced overlapping ROC curves throughout a FAR range of 0–0.16 ( $\sim 1/6$ ; the maximum possible FAR value for the six-person lineup)<sup>3</sup> for equivalent levels of discriminability. Using the same model and decision rule, we generated ROC curves from simulated data based on lineup sizes ranging from one to eight, for three levels of discriminability ( $d' = 0.5, 1$ , and  $1.5$ ). As shown in Figure 2, ROCs for lineups of varying sizes overlap throughout the FAR range of interest. That is, according to this very rudimentary model of eyewitness decision-making, there should be no meaningful effects of lineup size on discriminability.

### The Present Study

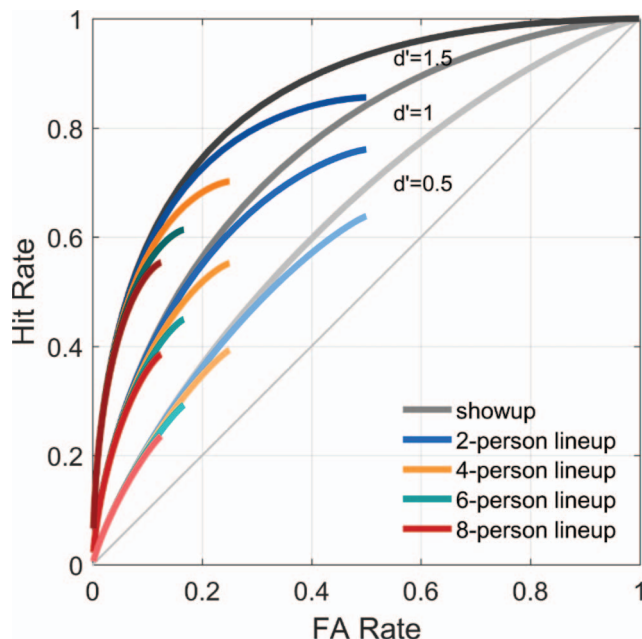
In the current study, we used ROC analyses to examine the effect of lineup size on diagnostic accuracy. We manipulated lineup size both as a between-subjects (Experiment 1) and a

within-subjects variable (Experiment 2). In the first experiment, participants viewed a video of a mock crime, followed by a lineup trial. In the second experiment, the targets were sequentially presented as still photographs and the test consisted of lineups of different sizes that either did or did not contain a target. In both experiments, participants were asked to provide a confidence rating following their responses.

In addition to direct examination of the ROCs that arise from different lineup sizes, we also fit competing variants of two signal-detection models to the data, namely the IO model and the *Ensemble* model, and assessed them by their ability to predict unseen data using cross-validation. These two models are currently the most prominent competing signal detection models of eyewitness memory and have been shown to provide good fits to empirical data (Wixted et al., 2018). As demonstrated in Figure 1, the IO model does not predict any changes in discriminability across showups and lineups of different sizes. The Ensemble model (Wixted et al., 2018) is derived from research in perception that demonstrates that the visual system quickly extracts summary representations from similar looking objects (e.g., Albrecht & Scholl, 2010), including faces (e.g., Neumann et al., 2013). According to the Ensemble model, then, the faces in a lineup are not evaluated independently of each other. Rather, each lineup member is compared with a summary representation extracted from the entire lineup. The exact means by which this occurs will become clear when we introduce the mathematical formulation of the models. The predictions of the Ensemble model on the effect of lineup size depend on the relationship between the memory signals elicited by each lineup member.

The IO model makes the assumption that there is no dependency between targets and fillers within a lineup in terms of their level of familiarity or memory strength. The Ensemble model, on the other hand, allows for the assumption of within-lineup dependencies (i.e., correlated memory signals). Within-lineup dependency simply refers to faces within a lineup being more similar to each other than to faces from different lineups, and thereby, generating correlated memory signals (see Wixted et al., 2018 for a thorough discussion). Put differently, the level of familiarity elicited by the target face is predictive of the level of familiarity that will be elicited by the fillers. Imagine a case in which an eyewitness describes the perpetrator as a short White male with thick eyebrows. A fair lineup that is created based on this description will have members—the suspect and the fillers alike—that all possess these characteristics. Thus, if the suspect generates a strong memory signal, the signals generated by the fillers should also be relatively strong, reducing the variance of memory strengths across members in a given lineup. Thus, a high correlation within lineups will give rise to low within-lineup variability ( $\sigma_w^2$ ) and high between-lineup variability ( $\sigma_b^2$ ). The correlation,  $\rho$ , or variance

**Figure 2**  
*Simulated ROC Curves From the Independent Observations MAX Model for the Showup, Two-, Four-, Six-, and Eight-Person Lineup Procedures When Discriminability was Equated Across the Procedures*



*Note.* The innocent and guilty suspects were randomly drawn from two normal distributions with equal variance. The innocent suspect and the fillers were drawn from a distribution with  $\mu = 0$  and  $\sigma = 1$ .  $\mu$  for the guilty suspect distribution ( $d'$ ) was set to 0.5, 1, or 1.5. Recall that the upper bound of the FAR is determined by lineup size such that the maximum FAR value is equal to  $1/n$ , where  $n$  represents the number of faces in a lineup. Therefore, the length of the receiver-operating characteristic (ROC) curves decreases with an increase in lineup size. FA = false alarm. See the online article for the color version of this figure.

<sup>3</sup> Both simulations demonstrate that when latent discriminability is identical across lineup size conditions, ROCs overlap initially; however, as one moves to the upper right in the ROC space, the ROC curves plot lower (i.e., become more concave-down) as lineup size increases. This is a consequence of the effect of lineup size on hit rates even when discriminability is fixed. Thus, when only more liberal response criteria are considered, the ROC curves for smaller lineups might lie slightly above the ROC curves for larger lineups for unimportant reasons (see Lampinen, 2016). This effect is shown here in Figure 2.

shared by members of a given lineup, is equal to the ratio of this between-lineup variability to the total variability ( $\rho = \sigma_b^2 / (\sigma_w^2 + \sigma_b^2)$ ).

To generate predictions from the Ensemble model, we simulated three different data sets with the correlation of memory signals between targets and fillers varying across data sets ( $\rho = 0$ ,  $\rho = 0.4$ , and  $\rho = 0.8$ ). The mean of the underlying memory strength distribution for the target was set to 1 ( $d' = 1$ ) for all lineup procedures in all three simulations. For convenience, the total variance ( $\sigma^2$ ) was set to 1; thus,  $\rho$  reduced to  $\sigma_b^2$ .

As shown in the left panel of Figure 3, when memory signals elicited by the targets and fillers are independent ( $\rho = 0$ ), as in the IO model, the Ensemble model predicts lower discriminability for the lineups compared with the showup.<sup>4</sup> When a correlation is introduced, the lineups mostly exhibit an advantage over showups. Thus, according to the Ensemble model, discriminability is expected to be hurt by poorly constructed lineups, in which fillers do not resemble the target. This prediction was confirmed by the finding that unfair lineups, in which the suspect was not matched to the fillers on a distinctive feature, were associated with lower discriminability than fair lineups (Colloff et al., 2016). Further, the Ensemble model predicts an increase in discriminability with an increase in lineup size with diminishing returns, with the magnitude of the increase in discriminability decreasing as lineup size increases.

We compared the Ensemble model with three variants of the IO model. The IO variants differed in the way in which performance in different lineup sizes was related to each other. To preview the results, even though data from the current experiments were qualitatively more consistent with the predictions of the Ensemble model than the IO model, further analysis of model characteristics yielded ambiguous results, and motivated the postulation of a new theoretical contender—the *Dependent Observations* (DO) model. We will discuss this issue in some detail in the Model Evaluation and General Discussion sections.

## Experiment 1

This experiment used a between-subjects manipulation of lineup procedure. Each participant made an identification from a single showup or a single lineup trial after viewing a video of a mock crime. Administering a single lineup to each participant is standard procedure in eyewitness memory research, presumably because it is akin to the task eyewitnesses face in the real world. Typically, an eyewitness is given a single lineup trial and is asked to identify a perpetrator associated with a single witnessed event. Despite the ecological validity of this approach, collecting a single data point from each participant leads to low statistical power with traditional sample sizes. Thus, the current experiment had a large sample size.

## Method

### Participants

To achieve 80% power to detect a true effect, a sample size of  $N \sim 700$  was required for each lineup size condition assuming a small effect size ( $d = .15$ ) using an independent-samples *t*-test. Because each subject only contributed a single data point, the strategy was to collect as much data as possible given resource

constraints. We ended up with sample sizes ranging from 881 to 993 for each condition. Participants ( $N = 4,401$ , 52% female) with a mean age of 25.91 ( $SD = 8.84$ , range = 10–72) were recruited from Amazon Mechanical Turk (MTurk) and from the University of California, San Diego (UCSD). The MTurk participants ( $N = 2,561$ ) received a modest payment and the UCSD participants ( $N = 1,840$ ) received course credit for their participation.<sup>5</sup> Forty-seven percent of participants self-identified as Asian/Indian, 37% as White, 7% as Latin/Hispanic, 3% as Black/African American, and 5% as Other/Unknown. Each lineup size by target presence condition had approximately 400 participants. See the bottom row in Table 1 for the exact sample sizes in each condition.

### Stimuli

A 26-s video of a mock shoplifting crime showed a perpetrator (an Asian man wearing loose fitting clothing and a baseball cap) walking down an aisle of a convenience store. He looked around nonchalantly at items, picked up an item, and then approached the camera, providing the viewer a close look at his face for several seconds. He then rounded the corner while discreetly placing the item into his back pocket and walking away. A mugshot style photo was taken of the actor wearing a different shirt (indistinguishable from those worn by the fillers).

A separate group of participants ( $N = 19$ ) watched the video and provided information about the suspect's appearance. They identified him as Asian or Hispanic, with dark hair and eyes, stocky to muscular in build, height ranging from 5'8 to 6'1, and with an average weight of 170 pounds. Based on those descriptors, 64 fillers were gathered from the online database of the Florida Department of Corrections. All of the backgrounds were made uniform and filters (e.g., Gaussian noise) were applied to degrade the suspect's photo so that the resolution visually matched the photo quality of the fillers. All lineup photos were presented simultaneously.

### Design

Lineup size and target presence were manipulated as between-subjects variables. Lineup size was manipulated at five levels. Each participant received either a showup or a two-, four-, six-, or eight-person lineup. The TP lineups consisted of the photo of the culprit presented with filler photos, and the TA lineups consisted only of the filler photos. No specific innocent suspect was designated for TA lineups. The position of the suspect was random in TP lineups. Fillers were randomly selected from a pool of 64 photos for each lineup trial/participant. Each participant was randomly assigned to one of the 10 conditions: target presence (two) crossed with lineup size (five).

### Procedure

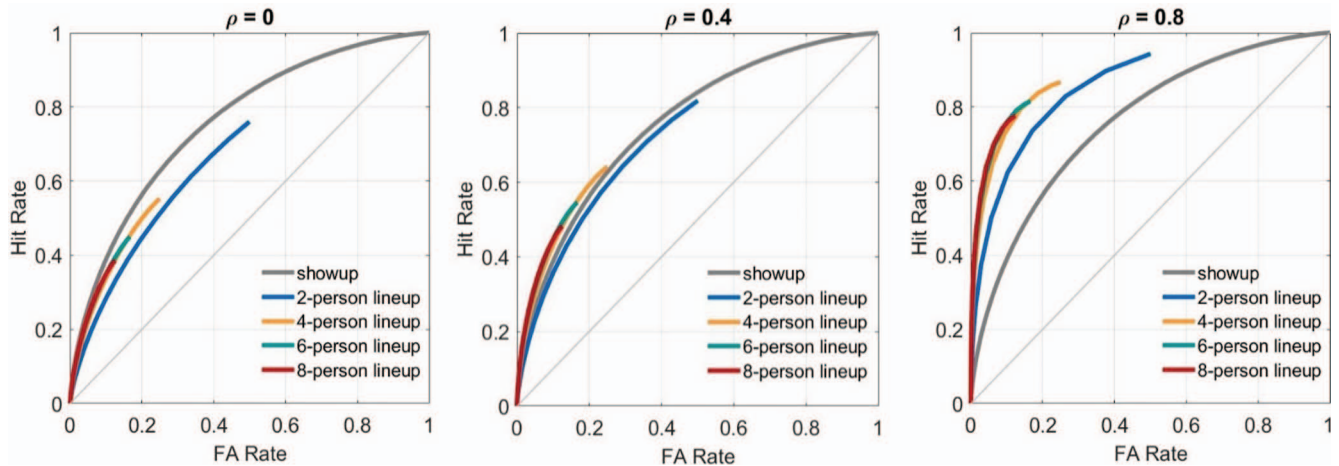
All participants completed the experiment online. Following a demographic survey, participants were told to pay close attention to the video. The showing of the video was followed by a 5-min distractor task (playing Tetris), which was immediately followed

<sup>4</sup> Note that the Ensemble model is not defined for a showup. The showup ROCs were based on the IO model.

<sup>5</sup> These data were collected and shared by Laura Mickes and John Wixted.

**Figure 3**

*Simulated ROC Curves for the Ensemble Model for the Showup, Two-, Four-, Six-, and Eight-person Lineup Procedures When Discriminability Was Equated Across the Procedures ( $d' = 1$ )*



*Note.* The innocent and guilty suspects were randomly drawn from two normal distributions with equal variance for the simulations, in which  $\rho = 0$  (left panel). For the other simulations ( $\rho = 0.4$  and  $\rho = 0.8$ ), the strength values of the innocent suspect and fillers were determined by the guilty suspect's memory strength. FA = false alarm; ROC = receiver-operating characteristic. See the online article for the color version of this figure.

by the lineup task. Participants were instructed that they would either see a showup or a lineup, and that the culprit from the video may or may not be in the lineup (or showup). The participants gave their response either by endorsing the showup or a lineup member, or by selecting the "not present" option. They indicated their confidence in their response on a decile scale ranging from 0% (*just guessing*) to 100% (*absolutely certain*). Multiples of 10 were the only possible ratings. Finally, participants were asked a few questions about the culprit from the video (e.g., "Did he have any tattoos?") and about the video ("What crime did the man commit?").<sup>6</sup>

## Results and Discussion

### Response Rates

The frequency counts for guilty and innocent suspect identifications at each level of confidence for each lineup size is shown in Table 1. The distribution of response types (target/filler/innocent identification, lineup rejection) for each lineup procedure is shown in Figure 4. As expected, both HR,  $\chi^2(4) = 74.00$ ,  $p < .0001$ , and FAR,  $\chi^2(4) = 51.27$ ,  $p < .0001$ , decreased as the lineup size increased. Comparison of overall hit and false alarm rates does not speak to the question of how lineup size affects discriminability, as reviewed earlier. One would expect to see a reduction in HR and FAR values even under the assumption that discriminability is identical for all conditions.

### ROC Analysis

Figure 5 shows the ROC curves for each lineup size. The points on the ROC curves correspond to the cumulated proportions of lineup trials, at each level of confidence, in which innocent suspect identifications (FAR) and guilty suspect identifications (HR) were made. ROC curves were fit using a binormal function, which fits a linear model to the standardized FAR and HR values over the

confidence bins. The innocent suspect ID rates (FARs) were estimated by dividing the filler identification rates in TA lineups by lineup size. The partial area under the curve (pAUC) was used as the summary statistic to compare lineup conditions. The reason for examining the partial area is that the hit and false alarm rates do not cumulate to 1 due to the omission of the filler identifications when constructing the ROC curve.

The pAUCs were computed over the truncated range of FAR values common to all conditions ( $0 < \text{FAR} < 0.065$ ). This range covers responses made using the entire confidence scale in the eight-person lineup condition and responses made with higher levels of confidence in other lineup size conditions. Because identification responses made with higher levels of confidence are more likely to be used as evidence in courts of law, the data that are in the upper-right hand quadrant of the ROC are less relevant for practical purposes. Additionally, considering the lower portion of the ROC handles the apparent small differences in the tail of ROC curves for lineup procedures with equivalent latent discriminability, observed in the more liberal regions of the ROC space (as shown in Figure 2).

For the statistical analyses comparing pAUC across conditions, we used the pROC package in R (Robin et al., 2011). The procedure involved computing the areas under the ROCs assuming linear functions between the data points (i.e., the trapezoidal method). Variability of pAUC was estimated by bootstrapping 10,000 samples from the original data set. See Figure 6 for pAUC values for each condition and their confidence intervals. The analyses revealed that pAUC for the showup (0.006) was smaller than both pAUC for the six-person lineup (0.013),  $D = 2.63$ ,  $p = .008$  and the eight-person lineup,  $D = 2.13$ ,  $p = .03$ . Other

<sup>6</sup> Analyses included data from all participants. Excluding participants who gave an incorrect answer to this question did not change the results. These data are made available on the project's OSF page.



**Table 1***Frequency Counts of Guilty Suspect and Innocent Suspect Identifications for Each Lineup Condition at Each Level of Confidence*

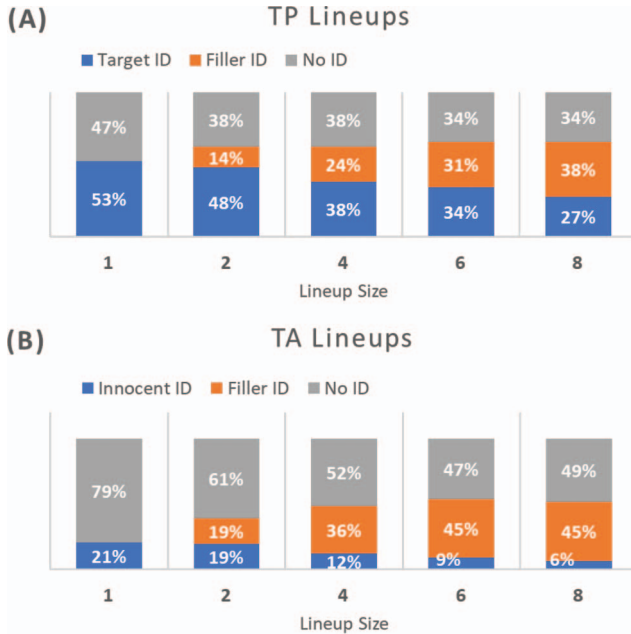
Confidence	Guilty suspect identification (TP)					Innocent suspect identification (TA)				
	Showup	Two-person lineup	Four-person lineup	Six-person lineup	Eight-person lineup	Showup	Two-person lineup	Four-person lineup	Six-person lineup	Eight-person lineup
0	7	5	3	1	2	13	7	2	3	1
10	4	3	2	1	1	6	3	2	1	1
20	8	3	4	1	0	3	5	3	2	1
30	20	13	8	3	3	6	9	6	3	3
40	30	7	10	8	8	13	7	6	5	3
50	30	32	21	24	13	15	16	8	6	4
60	38	18	24	21	22	7	9	8	6	4
70	28	35	39	26	24	4	17	7	5	5
80	30	20	18	34	15	14	8	7	4	4
90	28	22	25	26	16	4	3	2	2	2
100	16	42	17	14	9	9	8	5	2	1
0–100	239	200	171	159	113	94	177	209	224	228
<i>N</i>	455	417	448	462	411	458	454	432	419	445

*Note.* Number of innocent suspect IDs were estimated by dividing the raw frequency counts for each confidence level by lineup size. In cases where this estimate was a noninteger, it was rounded to the nearest integer.

pairwise comparisons did not yield statistically significant differences. Thus, consistent with prior work, we observed that showups led to lower discriminability between the guilty and the innocent suspect compared to (some) lineups.

**Figure 4**

*Response Rates as a Function of Lineup Size for TP and TA Lineups*



*Note.* (A) Proportions of target identifications, filler identifications, and no identifications in target present (TP) trials as a function of lineup size. (B) Proportions of innocent suspect identifications, filler identifications and no identifications in target absent (TA) trials as a function of lineup size. In this experiment, no innocent suspect was designated; the values for innocent identifications were computed by dividing the total number of filler identifications by the lineup size. All values were rounded to the nearest integer. See the online article for the color version of this figure.

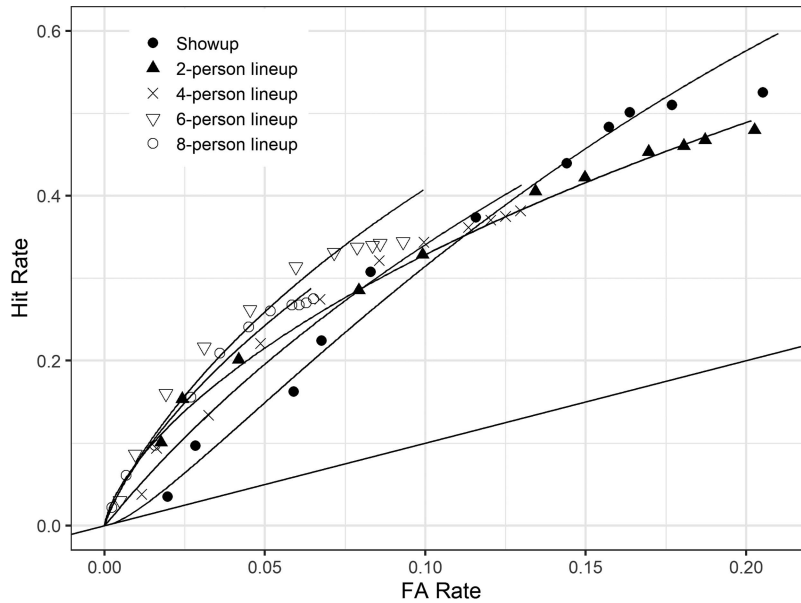
Differences in discriminability as a function of lineup size were not apparent. The only exception was the previously mentioned poorer performance evident for showups, replicating prior work (Gronlund et al., 2012; Mickes, 2015; Wetmore et al., 2017). The observed pattern, even though noisy, appears to be consistent with the predictions of the Ensemble model. The ROC curves for larger lineups tended to lie above those of smaller lineups up until the six-person lineup. This pattern was also evident in Wooten et al. (2020), who demonstrated an increase in discriminability from the showup to the three-person lineup procedure, as well as a numerical but nonsignificant increase in discriminability between three-person and six-person lineups, and no apparent differences in discriminability for lineups with more than six members. Experiment 2 assessed the generality of this finding by using a procedure that has much greater measurement precision on a per-subject basis.

For both this and the following experiment, we also examined the relationship between confidence and diagnostic accuracy by plotting confidence-accuracy calibration (CAC) functions for the different lineup sizes. These data are beyond the central purposes of our analysis and are presented in Appendix A.

## Experiment 2

In this experiment, lineup size was manipulated as a within-subjects variable in order to increase the precision of estimates for each participant's performance. Each participant completed multiple trials of each type of lineup procedure after viewing a list of target faces (perpetrators) as still photographs. As such, this procedure differed more substantially from a real-world eyewitness identification scenario. However, prior examination of this procedure has revealed that it has characteristics that generalize well to the study of eyewitness memory (Mansour et al., 2017). In that work, the multitrial nature of the procedure did not obscure or interact with the effects of key forensic variables such as lineup presentation (sequential vs. simultaneous) and memory strength on choosing rates from TA and TP lineups.

**Figure 5**  
ROC Curves for Each Lineup Size (Experiment 1)



*Note.* The points represent cumulative false alarm and hit rate pairs at each level of confidence. The leftmost points are based on identifications made with the highest level of confidence (100%), and the rightmost points are based on identifications made with any level of confidence, from 0% to 100%. The curves are fit using a binormal function. The diagonal line represents chance performance. FA = false alarm; ROC = receiver-operating characteristic.

## Method

### Participants

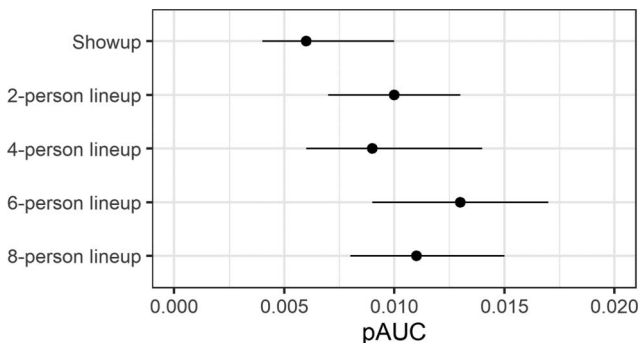
To achieve 80% power to detect a true effect using a paired-sample *t* test, a sample size of 90 was required assuming a small-to-medium effect size ( $d = .30$ ). We exceeded the aimed sample size thanks to the easy availability of subjects. Undergraduates ( $N = 105$ , 81% female) with a mean age of 19.22 ( $SD = 1.10$ , range = 18–22) from the University of Illinois at Urbana-

Champaign participated in return for course credit. Forty-three percent self-identified as Asian, 41% as White, 13% as Latin/Hispanic, 7% as Black/African American, and 9% as Other/Unknown. (The sum of the values exceeds 100% because participants were allowed to select more than one category.)

### Stimuli

Sixty sets of mugshots were gathered from the online database of the Florida Department of Corrections. Each set consisted of seven members who were matched on race, gender, age, hair color, hair style, eye color, height, and weight. See Figure 7 for a sample set. All mugshots belonged to people within the ages of 18 to 27. Half of the sets were composed of male members and the other half of the sets were composed of female members. Of the 60 sets of faces, 43 were composed of Whites, 10 were composed of Hispanics, and the remaining seven sets were composed of African Americans. This distribution approximately reflects the racial demographic profile of Illinois except for the lack of Asians in our stimuli set. This was due to the small number of Asians in the Florida Corrections Database. The backgrounds of all mugshots were rendered to be uniform in tone and all clothes were colored black. Distinctive elements such as tattoos, earrings, or makeup were removed from the photos. From each of the 60 sets, one member was randomly selected to serve as the guilty suspect (i.e., target), another member was randomly selected to serve as the innocent suspect, and the remaining five members were designated as fillers. The guilty and innocent suspects were kept constant across participants.

**Figure 6**  
Mean *pAUC* Values With 95% Bootstrapped Confidence Intervals (Experiment 1)



*Note.* *pAUC* = partial area under the curve.

**Figure 7**  
*An Example Set of Faces*



*Note.* A total of 60 sets were created and each set was randomly assigned to one of six conditions for each participant. All participants were shown the same set of 60 faces in the study phase—the target face from each set. For each set, there was a designated target (guilty suspect) and a designated innocent suspect that replaced the target in TA trials. Fillers were chosen randomly from the remaining five faces. See the online article for the color version of this figure.

**Design**

Lineup size (one-, three-, and six-person lineups) and target presence (TP vs. TA) were manipulated as within-subjects variables, yielding six within-subject conditions. For each condition, participants received 10 lineup trials, yielding a total of 60 trials (20 showups, 20 three-person lineups, and 20 six-person lineups). Half of the trials from each lineup condition included a target (TP lineups), and, for the other half of the lineup trials, the target was replaced with the innocent suspect (TA lineups). The assignment of mugshot sets to lineup conditions and the presentation order of the target pictures during study were randomized for each participant. The presentation order of the lineup trials was identical to the presentation order of the target faces in the study phase. This choice means that retention interval was more or less held constant across stimuli. In the TP lineups, a target was presented alone (showup condition) or with fillers other than the predesignated innocent suspect from the same set. For the three-person lineups, two fillers from the same set were randomly selected from five fillers, for each participant. The arrangement of photos in each lineup trial as well as the position of the suspect was random. The three-person lineups were displayed on the screen as a 1 × 3 image array, and the six-person lineups were displayed as a 2 × 3 image array. No stimulus set was used more than once for each participant.

**Procedure**

After completing a demographic survey, participants were instructed that they would be shown a series of faces one by one and

were asked to study them carefully for a later memory test. They were not informed of the number of faces they would study or of the type of the subsequent memory test. Sixty faces—the target photo from each set—were presented sequentially, each for 4 s, with a 0.75-s interstimulus interval. When the study phase was over, participants engaged in a distractor task in which they solved simple arithmetic problems for 2.5 min. Participants were then given instructions for the lineup task. They were informed that on each trial, they would see one, three, or six faces from which they would need to choose the one that they studied earlier or to choose the “absent” option if they did not remember seeing any of the lineup members. They were also informed that none of the lineups would contain more than one previously studied face. They then proceeded to 60 self-paced lineup trials. After making their response, participants were asked, “How confident are you that the face you selected is the one you studied?” or if they rejected the lineup, they were asked, “How confident are you that you did not study any of the faces in the lineup?” They responded using a three-level confidence scale with the following options: “I am just guessing,” “I think I am right,” and “I am sure I am right.”

**Results and Discussion**

**Response Rates**

See Table 2 for the frequency counts of target (guilty suspect) identifications and innocent suspect identifications for each level of confidence and collapsed over all levels of confidence. Figure 8 shows mean response proportions across conditions.

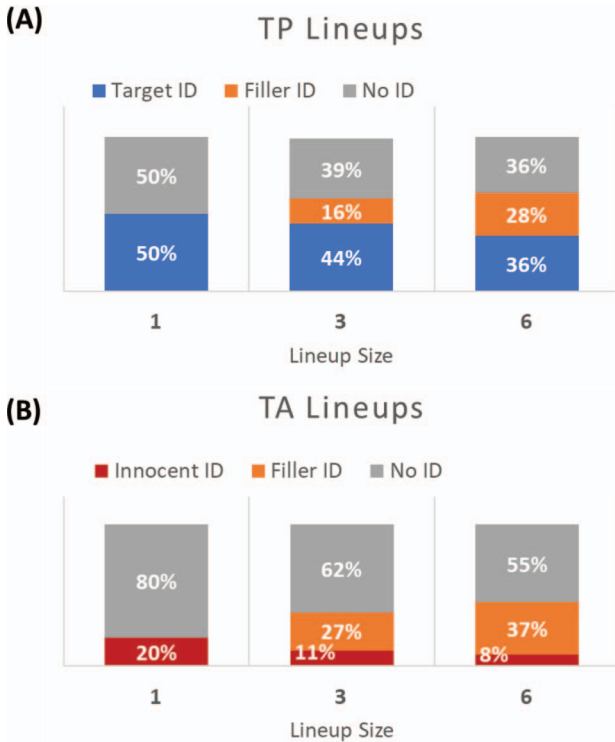
**Table 2**  
*Number of Trials in Which a Guilty Suspect Was Identified in Target Present (TP) Lineups and Number of Trials in Which the Designated Innocent Suspect Was Identified in Target Absent (TA) Lineups for Each Level of Confidence Collapsed Across All Participants*

Confidence	Guilty suspect identification (TP Trials)			Innocent suspect identification (TA Trials)		
	Showup	Three-person lineup	Six-person lineup	Showup	Three-person lineup	Six-person lineup
Low (1)	65	60	31	49	31	24
Medium (2)	210	189	129	120	66	49
High (3)	251	217	218	37	20	11
Overall	526	466	378	206	117	84

*Note.* The total number of trials was 1,050 (10 trials from each of 105 participants) for each lineup Size × Target Presence condition.

**Figure 8**

Response Rates as a Function of Lineup Size for TP and TA Lineups



*Note.* (A) Proportions of target identifications, filler identifications, and no identifications (responding “absent”) in target present trials as a function of lineup size. (B) Proportions of innocent suspect identifications, filler identifications, and no identifications in target absent trials as a function of lineup size. For each condition, the proportions are based on a total of 1,050 trials (10 trials from 105 participants). TP = target present; TA = target absent. See the online article for the color version of this figure.

As expected, hit rates and false-alarm rates both decreased with increasing lineup size. Lineup size was a significant predictor of hit rates,  $\chi^2(2) = 34.75, p < .0001$ . The hit rate was significantly lower in the three-person lineup than in the showup,  $t(104) = 2.57, p < .05$ , and in the six-person lineup than in the three-person lineup,  $t(104) = 3.45, p < .001$ . Lineup size had a similar effect on false alarm rates,  $\chi^2(2) = 49.59, p < .0001$ . Participants were less likely to choose the innocent suspect from the three-person lineup than in the showup,  $t(104) = 4.91, p < .0001$ , and from the six-person lineup than the three-person lineup,  $t(104) = 2.64, p < .01$ . In addition, lineup rejections also decreased with increasing lineup size for both TA lineups,  $\chi^2(2) = 116.94, p < .0001$ , and TP lineups,  $\chi^2(2) = 34.94, p < .0001$ . Participants were less likely to reject a three-person TA lineup than a TA showup,  $t(104) = 8.39, p < .0001$ , and less likely to reject a three-person TA lineup than a six-person TA lineup,  $t(104) = 4.06, p < .0001$ . Similar to TA lineups, participants were less likely to reject a three-person TP lineup than a TP showup,  $t(104) = 4.85, p < .0001$ , but the reduction in rejection rates from three-person to six-person TP lineups was not statistically significant,  $t(104) = 1.39, p > .05$ .

The same pattern is evident within each confidence category, as shown in Table 2.

### ROC Analysis

Figure 9 shows the ROC curves for each lineup condition. The points on the ROC curves are the group FAR and HR values based on responses made at each cumulated confidence category. The procedure for constructing the ROC curve was identical to Experiment 1. The pAUCs were estimated over the truncated range of FAR values that were common to all conditions ( $0 < \text{FAR} < 0.08$ ).

This analysis revealed that the pAUC for the showup (0.016, 95% CI [0.014, 0.019]) was significantly lower than the pAUC for the six-person lineup (0.021, 95% CI [0.019, 0.024]),  $D = 3.01, p = .003$ . The difference between the pAUCs for the showup and the three-person lineup (0.020, 95% CI [0.017, 0.023]) was not significant,  $D = 1.87, p = .06$ , though the effect size was suggestive and indicates a need for higher-powered studies. The difference between the three- and the six-person lineup was not significant ( $D = .30, p = .76$ ). See Figure 10 for the pAUC values for each condition and their confidence intervals.

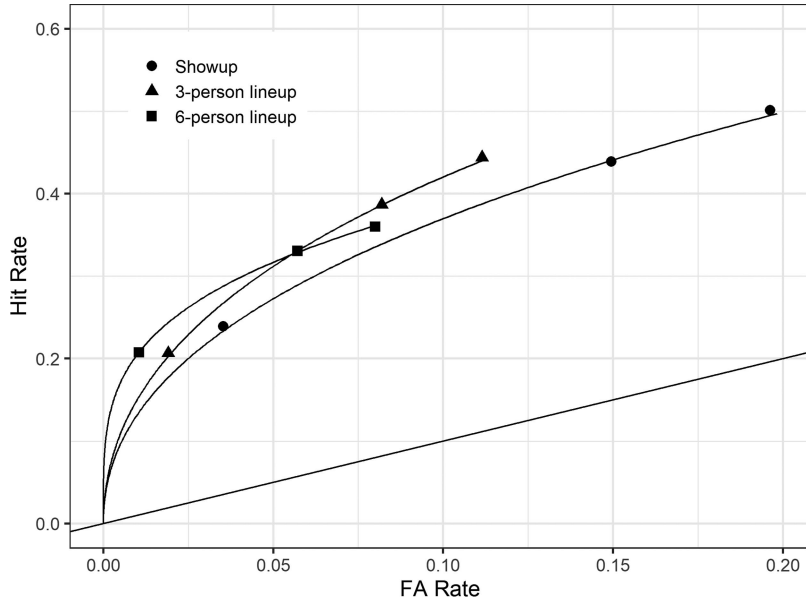
The results from this experiment replicate Experiment 1, as well as findings from prior work indicating poorer diagnostic performance of showups compared with lineups (Gronlund et al., 2012; Mickes, 2015; Wetmore et al., 2017). It again did not reveal any differences in discriminability between lineups of greater than one face. However, the pattern that was observed in Experiment 1 was also apparent here. That is, there was a numerical increase in discriminability with increasing lineup size, a result that is ordinarily consistent with the predictions of the Ensemble model.

### Model Comparison and Evaluation

Our approach to modeling eyewitness recognition is based on SDT. We assume that recognition requires an eyewitness to make a decision based on evidence that is impacted by random variation, or noise (Wickens, 2002). For ease of explication, we begin by describing this decision process in a showup procedure, in which the eyewitness must judge whether an innocent or guilty suspect matches their memory of the perpetrator. In this situation, the eyewitness considers the amount of evidence that a guilty or innocent suspect elicits in matching their memory of the perpetrator. Because memory representations are inexact, the amount of evidence generated by the suspect will be variable. Consistent with typical signal detection interpretations of eyewitness performance (e.g., Gronlund & Benjamin, 2018; Wixted & Mickes, 2018), we assume Gaussian distributions of noise around a true value, and assume with no loss of generalizability that the distribution of evidence generated by an innocent suspect follows a normal distribution centered at zero, with a standard deviation equal to one. The amount of evidence generated by the guilty suspect is also assumed to be normally distributed; however, the mean of this distribution is assumed to be greater than that of the distribution of evidence generated by the innocent suspect. This captures the fact that, from the perspective of the eyewitness, the guilty suspect should typically generate more evidence because he or she is actually the perpetrator that the eyewitness saw. Because each person generates some amount of evidence, the eyewitness must set a decision criterion for responding whether a suspect is the perpetrator or not.



**Figure 9**  
ROC Plots for Each Lineup Size (Experiment 2)



*Note.* The points represent false alarm and hit rate pairs at each cumulated level of confidence. The leftmost points refer to identifications made with the highest level of confidence (most conservative response criterion). The curves are fit using a binormal function. The diagonal line represents chance performance. ROC = receiver-operating characteristic.

Next, we extend this signal detection logic to the lineup procedure in which the eyewitness must identify the perpetrator in a lineup of faces. A lineup contains the suspect and  $n-1$  fillers, where  $n$  denotes the total number of people in the lineup. Unlike the real world, we know in our experiments whether the suspect is truly guilty or innocent.

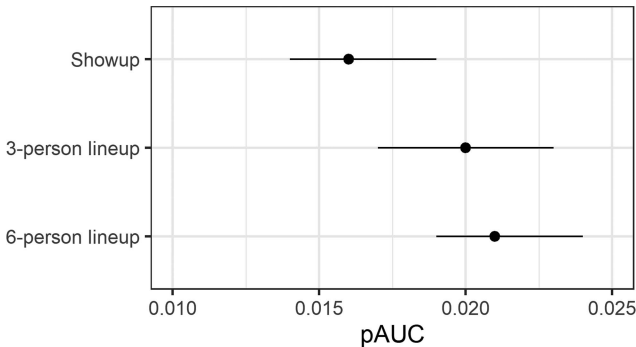
We start by considering two signal detection models of eyewitness identifications from lineups: the IO model and the Ensemble model. Both models are based on a MAX decision rule, such that the face that generates the strongest evidence for matching the memory of the perpetrator is identified if that amount of evidence

exceeds the eyewitness's decision criterion. The two models employ different assumptions about the nature of how evidence is gathered from a face. For the IO model, identification decisions in a lineup are simply based on the untransformed strength value of the signal generated by the face with the strongest memory signal. Strength values for the individual faces are assumed by fiat to be uncorrelated. The diagnostic decision variable in the Ensemble model, on the other hand, is the difference between this strongest memory signal and the average of the memory signals generated by all members in the lineup.

### Independent Observations Model

Here we provide a mathematical description of the MAX rule within the context of the IO model to derive predictions about hit and false alarm rates. We use  $\varphi\left(\frac{x-\mu_{GS}}{\sigma_{GS}}\right)$  and  $\Phi\left(\frac{x-\mu_{GS}}{\sigma_{GS}}\right)$  to denote the probability and cumulative density function representing the amount of evidence for a memory match generated by the guilty suspect, respectively. We use  $\varphi(x)$  and  $\Phi(x)$  to denote the probability and cumulative density function that represents evidence for a memory match generated by the innocent suspect or by fillers in a lineup, respectively. It is generally assumed that in a fair lineup, the characteristics of the fillers are matched to the description of the suspect and that from the standpoint of the eyewitness, an innocent suspect should generate a similar memory signal to the remaining fillers (although see Clark, 2003 for explanation of conditions under which this assumption might not hold). Thus, we denote the innocent suspect and filler distributions identically.

**Figure 10**  
Mean  $pAUC$  Values and 95% Bootstrapped Confidence Intervals (Experiment 2)



*Note.*  $pAUC$  = partial area under the curve.

In a showup procedure the predicted hit rate is simply the probability that the memory signal generated by the guilty suspect (target) exceeds the decision criterion. In a TP lineup procedure with  $n$  members (one target and  $n-1$  fillers), the predicted hit rate is based on the probability that the memory signal ( $x$ ) generated by the guilty suspect exceeds the decision criterion as well as all of the memory signals generated by the fillers in the lineup. Formally, the hit rate predicted by the model for both the showup and the lineup procedure is given by the following equation:

$$p(\text{Target ID} | TP) = \int_{-\infty}^{\infty} \varphi\left(\frac{x - \mu_{GS}}{\sigma_{GS}}\right) \times \phi^{n-1}(x) \times \left(1 - \phi\left(\frac{C-x}{\sigma_C}\right)\right) dx$$

where  $C$  is the decision criterion and  $n$  is the number of faces in the lineup. Note that the variability in criterion placement represented by  $\sigma_C$  is set to zero because we assume no criterion variability.

A similar logic applies to the prediction of false alarm rates. In this case, the predicted false alarm rate is based on the endorsement of a single innocent suspect. In the showup procedure, the false alarm rate is simply the probability that the signal generated by the innocent suspect exceeds the decision criterion. In the lineup procedure, the false alarm rate is the probability that the signal generated by the innocent suspect (one of the fillers) exceeds the decision criterion as well as the memory signals produced by the remaining  $n-1$  fillers in the lineup:

$$p(\text{Innocent suspect ID} | TA) = \int_{-\infty}^{\infty} \varphi(x) \times \phi^{n-1}(x) \times \left(1 - \phi\left(\frac{C-x}{\sigma_C}\right)\right) dx$$

The filler identification rate in a TA lineup in which there is no designated suspect is given by multiplying the equation for innocent suspect IDs by lineup size:

$$p(\text{Filler ID} | TA) = n \times \int_{-\infty}^{\infty} \varphi(x) \times \phi^{n-1}(x) \times \left(1 - \phi\left(\frac{C-x}{\sigma_C}\right)\right) dx$$

Recall that for TA lineups, when an innocent suspect is not designated, the false alarm rate is estimated by dividing the probability of filler identifications by the size of the lineup. This returns an estimate of the probability of an innocent suspect identification based on the assumption that the innocent suspect is just another filler from the eyewitness's perspective. We also considered the situation in which one of the fillers is incorrectly identified as the perpetrator in a TP lineup. We refer to these kinds of responses as filler identifications that are conditional on the guilty suspect being present in the lineup. In the showup procedure, the probability of this event is zero, since no fillers are shown along with the guilty suspect. In a lineup procedure, the probability of this event is the joint probability that the memory signal generated by any one of the fillers exceeds the signal generated by the guilty suspect, the  $n-2$  remaining fillers, and the decision criterion. Thus, the predicted probability of filler identification in a lineup when the guilty suspect is present is given by the following equation:

$$p(\text{Filler ID} | TP) = (n-1) \int_{-\infty}^{\infty} \varphi(x) \times \phi\left(\frac{x - \mu_{GS}}{\sigma_{GS}}\right) \times \phi^{n-2}(x) \times \left(1 - \phi\left(\frac{C-x}{\sigma_C}\right)\right) dx.$$

Finally, in order to apply our model fitting procedure, we also considered cases in which no suspect or filler was identified in either the showup or lineup procedures. On TA trials, rejections of a lineup are equivalent to the complement of the probability that any of the fillers are identified. The predicted probability for rejections on target-absent trials is thus given by the following equation:

$$p(\text{Lineup rejection} | TA) = 1 - p(\text{Filler ID} | TA).$$

On TP trials, rejections of a lineup are equivalent to the complement of the probability that either the guilty suspect is endorsed or one of the fillers is endorsed. The predicted probability for rejections on target present trials is given by the following equation:

$$p(\text{Lineup rejection} | TP) = 1 - [p(\text{Filler ID} | TP) + p(\text{Target ID} | TP)].$$

## Ensemble Model

According to the Ensemble model (Wixted et al., 2018), the eyewitness takes the difference between the memory signal generated by each face in the lineup and the average of the memory signal generated by all faces in the lineup. Then, in a manner identical to that built into the IO model, the eyewitness's response strategy is based on the MAX rule. The eyewitness identifies a face from a lineup if and only if the difference between the memory signal associated with that face and the average memory signal from the lineup (a) is larger than the differences between the memory signals associated with each individual face and the average memory signal from the lineup and (b) exceeds the decision criterion. In situations where the maximum difference between each face's memory signal and the average of the memory signals from all faces in the lineup does not exceed the decision criterion, the observer will respond "perpetrator absent."

The Ensemble model is not defined for the showup procedure because the mean strength value of a single face is the same as the actual strength value for that face. For drawing predictions from the Ensemble model, we set the predicted hit rate to be the probability that the memory signal generated by the guilty suspect (target) exceeds the decision criterion, exactly as in the IO model. In a TP lineup procedure with  $n$  members, the predicted hit rate is based on the probability that the difference between the memory signal generated by the guilty suspect ( $x$ ) and the average of the memory signals generated by all members in the lineup exceeds all of the differences between memory signals generated by the  $n-1$  fillers in the lineup and the average memory signal (i.e., the guilty suspect generates the maximum signal) as well as the decision criterion. Formally, the hit rate predicted by the model for both the showup and lineup procedures is given by the following equation:

$$p(\text{Target ID} | TP) = \int_{-\infty}^{\infty} \varphi\left(\frac{x - \mu_{GS}}{\sigma_{GS}}\right) \times \phi^{n-1}(x) \times \left(1 - \phi\left(\frac{C - (x \times (1 - 1/n) - (n-1) \times \mu_{TR,F})}{\sqrt{(n-1) \times \sigma_{TR,F}^2}}\right)\right) dx$$

Unlike the IO model, the Ensemble model assumes that a positive ID is made only if the difference between the memory signal  $x$  and the average memory signal from the lineup exceeds the decision criterion  $C$ . This is represented by the last term in the right-hand

side of the above equation. The terms  $\mu_{TR,F}$  and  $\sigma_{TR,F}$  denote the truncated (represented by the letters TR in the subscripts) Gaussian approximation of the mean and variance of each filler (represented by the letter F in the subscripts) in the lineup, respectively. The upper bound for the memory signal of each filler is equal to the memory signal  $x$ , capturing the fact that, in this scenario, the memory signal of each filler in the lineup must be smaller than  $x$ —the memory signal generated by the guilty suspect. For a complete derivation of the model's likelihood, see Wixted et al. (2018).

A similar logic applies to the prediction of false alarm rates. In this case, the predicted false alarm rate is based on the endorsement of a single innocent suspect. In the showup procedure, the false alarm rate is simply the probability that the signal generated by the innocent suspect exceeds the decision criterion. In the lineup procedure, the false alarm rate is the joint probability that the signal  $x$  generated by the innocent suspect (one of the fillers) exceeds the memory signals produced by the remaining  $n-1$  fillers in the lineup, and the probability that difference between the memory signal generated by the innocent suspect and the average memory signal from the lineup exceeds the decision criterion:

$$p(\text{Innocent suspect ID} | TA) = \int_{-\infty}^{\infty} \varphi(x) \times \phi^{n-1}(x) \times \left( 1 - \phi \left( \frac{C - (x \times (1 - 1/n) - (n-1) \times \mu_{TR,F})}{\sqrt{(n-1) \times \sigma_{TR,F}^2}} \right) \right) dx$$

The filler identification rate in a TA lineup in which there is no designated suspect is given by multiplying the equation for innocent suspect IDs by the lineup size:

$$p(\text{Filler ID} | TA) = n \times \int_{-\infty}^{\infty} \varphi(x) \times \phi^{n-1}(x) \times \left( 1 - \phi \left( \frac{C - (x \times (1 - 1/n) - (n-1) \times \mu_{TR,F})}{\sqrt{(n-1) \times \sigma_{TR,F}^2}} \right) \right) dx$$

As with the IO model, we considered the situation in which one of the fillers is incorrectly identified as the perpetrator in a TP lineup. In the showup procedure, the probability of this event is zero, because no fillers are shown along with the guilty suspect. In a lineup procedure, the probability of this event is the joint probability that the memory signal generated by any one of the fillers exceeds the signal generated by the guilty suspect, the  $n-2$  remaining fillers, and that the difference between the memory signal generated by one of the fillers and the average memory signal generated by the lineup exceeds the decision criterion. Thus, the predicted probability of filler identification in a lineup when the guilty suspect is present is given the by the following equation:

$$p(\text{Filler ID} | TP) = (n-1) \times \int_{-\infty}^{\infty} \varphi(x) \times \phi \left( \frac{x - \mu_{GS}}{\sigma_{GS}} \right) \times \phi^{n-2}(x) \times \left( 1 - \phi \left( \frac{C - (x \times (1 - 1/n) - ((n-2) \times \mu_{TR,F} + \mu_{TR,GS}))}{\sqrt{(n-2) \times \sigma_{TR,F}^2 + \sigma_{TR,GS}^2}} \right) \right) dx$$

In the above equation  $\mu_{TR,GS}$  and  $\sigma_{TR,GS}^2$  are the Gaussian approximations of the mean and variance, respectively, of the conditional distribution of memory signals generated by the guilty suspect, which is truncated at  $x$ , the maximum memory signal generated by one of the fillers.

Finally, as before, we use the probability of lineup rejections on TP and TA trials to implement the modeling. These expressions are identical to those used for the IO model.

## Evaluated Models

We tested the Ensemble model and three variants of the IO model, with each variant embodying different predictions about the relationship between lineup size and an eyewitness's ability to discriminate the memory signal generated by the guilty suspect from that generated by the innocent suspect. When fitting our models we assumed that the eyewitness's ability to discriminate the memory signal in a given experimental procedure is governed by two latent variables, which are modeled by the (a) distance between the distribution of the evidence generated by the fillers (one of which is the innocent suspect) and the distribution of evidence generated by the guilty suspect ( $\mu_{GS}$ ); and (b) the variance of the distribution of evidence generated by the guilty suspect ( $\sigma_{GS}$ ). We used confidence ratings to estimate decision criteria used when making an identification decision. We made no predictions regarding how lineup size affects decision criteria; therefore, the criterion parameters ( $c$ ) were free to vary across confidence bins and lineup sizes for all the evaluated models.

### Ensemble Model

Given fixed latent variables,  $\mu_{GS}$  and  $\sigma_{GS}$ , across varying lineup sizes, the Ensemble model predicts an increase in manifest discriminability with increasing lineup size (including the showup), with the rate of increase decelerating as lineup size increases. Notably, this pattern—apparent both in Experiment 1 and 2—is an inherent prediction of the Ensemble model only when there is a high correlation between the strengths of the memory signals elicited by the targets and the fillers. This is an important point that we will revisit shortly. As shown in the left panel of Figure 3, the Ensemble model also predicts that an eyewitness will gather lower-quality evidence from the lineup compared with the showup procedure, if the lineup is constructed poorly (i.e., the correlation between target and filler memory strengths are low), leading to a higher manifest discriminability in the showup condition.

### Fixed Latent Variable IO Model

The *Fixed* variant of the IO model makes assumptions consistent with the pure IO model, namely that lineup size has no differential effect on the eyewitness's ability to discriminate the perpetrator from an innocent suspect and that performance across the showup procedure and different lineup procedures is governed by two common latent variables (i.e.,  $\mu_{GS}$  and  $\sigma_{GS}$ ). We refer to this model as the *Fixed Latent variable* IO model because parameters used to estimate latent variables are fixed across different lineup sizes.

### Free Latent Variable IO Model

This *Free* variant of the IO model assumes that lineup size has an effect on discriminability but does not specify the nature of this effect. In this sense, this model is the most flexible; it assumes that there are  $2 \times n_L$  latent variables ( $\mu_{GS}$  and  $\sigma_{GS}$ ) that govern the eyewitness's ability to discriminate the guilty suspect from the innocent suspect across different lineup sizes, where  $n_L$  denotes the number of different lineup sizes. We refer to this model as the

*Free Latent variable* IO model because parameters used to estimate the latent variables across different conditions are free to vary. This is the model in which no meaningful relationships between performance on different lineup sizes are assumed.

### Hybrid IO Model With Free Showup Condition

We considered a third IO model that makes a prediction that is intermediate to the Fixed and Free latent variable models. This model assumes that performance in the lineup procedure differs from performance in the showup procedure, but that performance in the lineup procedure does not change with changes in lineup size. We refer to this model as a *Hybrid* model because one set of parameters is used to estimate latent variables in the showup condition, and a second set of parameters is used to estimate latent variables in different lineup conditions. This model was motivated by previous and current work showing a reduction in discriminability in showups compared to lineups (Gronlund et al., 2012; Mickes, 2015; Wetmore et al., 2017).

### Parameter Estimation and Model Comparison

We used within-sample split-half cross-validation to evaluate the models. Cross-validation analysis is a powerful model comparison technique that provides insight into how well a model generalizes to a new sample of data, rather than how well a given model fits a specific sample of data. The problem with guiding model selection based on model fit is that models may *overfit* a sample of data, that is, fit error variance instead of variance that is produced by a latent variable of interest (Yarkoni & Westfall, 2017). A direct way of assessing how well each model captures variance that is generated by latent variables is by directly testing how it generalizes to new samples of data. This is because models that erroneously fit error variance (overfit the data) within one sample should perform worse when used to predict a different sample of data where the distribution of noise differs (Robinson et al., 2020).

The analyses were implemented in MATLAB. The structure of the code was adapted from the ROC Toolbox (Koen et al., 2016) and specific model fitting and cross-validation algorithms were custom written. We used split-half cross-validation<sup>7</sup> to assess the various models. In both data sets (Experiments 1 and 2), each data point was treated as an independent observation. For each implementation, the data were randomly shuffled and split in half. The data were split separately within each lineup condition. Half of the data were used to construct points on an ROC curve (i.e., the training set). These averaged ROCs were then used to train the model—that is, to estimate best fitting parameters. We estimated best-fitting parameters by minimizing the negative log likelihood. Estimates of the best-fitting parameters from the training set were used to predict the data in the held-out half of the data (i.e., the validation set). This procedure was repeated 100 times with different random splits of the data, which yielded a total of 100 iterations for each implementation of cross-validation.

The critical outcome measure that we used to guide model selection was the mean squared error of cross-validation ( $MSE_{CV}$ ).  $MSE_{CV}$  was computed by averaging the sum of squared errors of cross-validation ( $SSE_{CV}$ , that is, the residual variance of prediction in the validation set across iterations ( $N=100$ ) of the cross-validation analysis.  $MSE_{CV}$  captures the amount of variance each

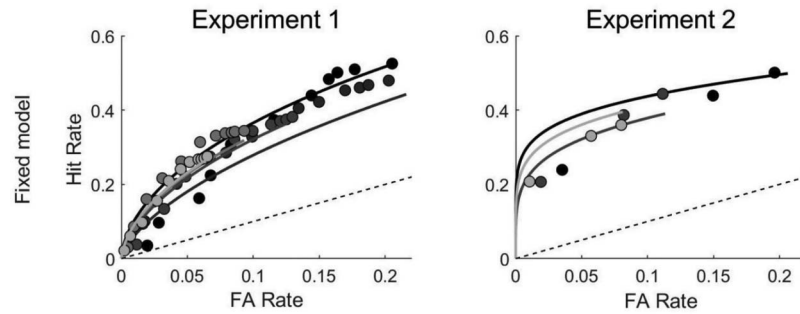
model failed to predict in the validation set. Models that yield relatively smaller values of  $MSE_{CV}$  perform better in terms of generalizing to a new sample of data. We indexed  $MSE_{CV}$  for each model using the standard deviation of  $SSE_{CV}$  across cross-validation iterations, with differences of greater than one standard deviation indicating clear support for the superior model. In addition, we compared each model nonparametrically by tallying the number of times that the  $SSE_{CV}$  of each model across the 100 cross-validation iterations was less than the  $SSE_{CV}$  of a competing model. We divided this value by 100 to calculate the proportion ( $P_{CV}$ ) of cross-validation iterations that the  $SSE_{CV}$  of a given model was smaller than the  $SSE_{CV}$  of a competing model. A value of  $P_{CV}$  close to .5 indicates that each of two models performed comparably in terms of prediction, whereas a value of  $P_{CV}$  close to 1 (or to 0) favors one of the two competing models.

Figure 11 and Figure 12 show the fits of the Ensemble model and the three variants of the IO model to all the data in Experiments 1 and 2. For completeness, in Table 3, we report traditional fit statistics calculated based on each model's fit to the entire set of data for each experiment along with the  $MSE_{CV}$ . Table 4 reports the best-fitting parameters. As shown, AIC consistently favored the Free latent variable IO model. In contrast, BIC favored the Fixed latent variable IO model consistently across both experiments. So, AIC favored the most flexible model (i.e., the model with the highest number of parameters) while the BIC favored the simplest model. Our use of cross-validation is intended to overcome the limitations of the AIC and BIC for models that differ in flexibility. We report AIC and BIC for archival value but base our conclusions on the results from the cross-validation analysis.

The results of the cross-validation analyses for both experiments are summarized in Table 3. As shown, the  $MSE_{CV}$  of all models were within one standard deviation of one another for both experiments. As such, the results from this analysis were not definitive. To further examine whether the empirical data were described better by a particular model, we compared the results for each pair of submodels nonparametrically. This procedure yielded six pairwise comparisons: the comparison of the Ensemble model with each of the three variants of the IO model, and three pairwise comparisons of the IO variants with each other. For each comparison, we report the proportion of iterations ( $P_{CV}$ ) in which one model was favored over the other (i.e., yielded a lower  $SSE_{CV}$ ) in Table 5. In both experiments, all three IO variants were favored over the Ensemble model, and both the Fixed and Hybrid IO models were favored over the Free IO model. In Experiment 1, the Hybrid IO performed better than the Fixed IO model, making the Hybrid IO the winning model. In Experiment 2, however, there was not a clear winner between the Hybrid and Fixed IO models, with the Hybrid model producing smaller error values 49% of the time. These outcomes align partly with the pattern evident in the

<sup>7</sup> We also assessed the models using five- and 10-fold cross-validation to examine how robust cross-validation results were when different sample sizes were used to estimate model parameters as well as to test the models. The pattern of results was the same across these different fold sizes. However, we found that using an unequal number of observations in the training and test set increased the discrepancy between the proportions generated in the training and test data sets, and consequently increased the overall mean squared error of cross-validation. Because the pattern of results was qualitatively the same, for ease of exposition we report only the results of the split-half cross-validation analysis.



**Figure 11***Empirical ROCs and the Best-Fitting ROC Curves for the Ensemble Model*

*Note.* Points denote the empirical data, and curves denote the model prediction based on best fitting values of  $\mu$  and  $\sigma$  for each model. Data and model fits in the showup procedure and the two-, four-, six-, and eight-person lineup procedures in Experiment 1 are denoted by black, dark gray, light gray, ash, and abalone shades, respectively. Data and model fits in the showup procedure and the three- and six-person lineup procedures in Experiment 2 are denoted by black, ash, and abalone shades, respectively. The graphs represent only the truncated ROC curves, that is, the proportion of guilty suspect identifications on target present trials and filler identification on target absent trials corrected by lineup size (Experiment 1) or innocent suspect identifications on target absent trials (Experiment 2). FA = false alarm; ROC = receiver-operating characteristic.

empirical ROC analyses from both experiments. We found support for the Hybrid model over the other models in Experiment 1, and partial support for that model in Experiment 2, where it proved to be superior to the Ensemble and Free IO models but on par with the Fixed IO model. Overall, the Hybrid model received the most support, indicating that showups elicit significantly lower latent discriminability, but the lineups varying in size do not differ significantly from one another. However, it is worth noting that this result was not fully replicated in the analysis of Experiment 2 and should be considered provisional until more data are brought to bear on the question.

### Model Evaluation Based on Data From Wooten et al. (2020)

The fact that the results were descriptively more in line with the predictions of the Ensemble model, but did not support that model, was surprising. We considered the possibility that the Ensemble model performed more poorly than the IO model because we included the showup condition when modeling the data. Including the showup condition may be problematic for evaluating the Ensemble model because the Ensemble model is not defined for the showup. Furthermore, the correlation between the target and filler strengths is not incorporated into the likelihood function of the Ensemble model. This is important because the predictions of the Ensemble model pertaining to the comparison of discriminability between showups and lineups depend on the magnitude of this correlation. This was demonstrated in Figure 3—an increase in correlation reverses the relationship between the showup and lineup procedures: The lineup is a superior procedure than the showup when this correlation is high, but is inferior to the showup when the correlation is nil. It should be noted that the correlation also affects the magnitude of the difference in discriminability between lineups; however, the ordinal rankings of different lineup

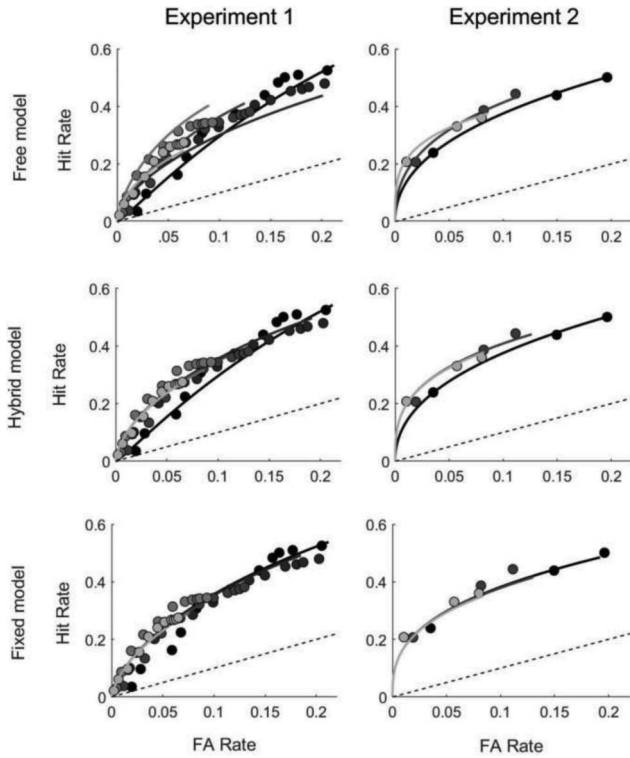
procedures stay the same—larger lineups elicit higher discriminability before leveling out at asymptote. For these reasons, we predicted that the Ensemble model might outperform the IO model when the showup condition is excluded. To examine this possibility, we evaluated each of the models using data from Wooten et al. (2020), who found a similar pattern of results to the present study, with a larger sample size. Because the showup condition was eliminated, we included only the Fixed and Free variants of the IO model along with the Ensemble model.

In Wooten et al. (2020), participants gave a confidence rating ranging on a scale from 0% to 100%. We collapsed across confidence ratings to generate four confidence bins. The particular cut-offs for each confidence bin were determined to obtain a sufficient number of data points for cross-validation. The cut-off values for each confidence bin were: 0–49%, 50–69%, 70–89%, and 90–100%. We evaluated the models using data only from the lineup conditions (three-, six-, nine-, and 12-person lineups) using the same cross-validation procedure. We report mean squared error of cross-validation ( $MSE_{CV}$ ) along with the fit statistics, AIC and BIC, for each model in Table 6. We again base our conclusions on the results of the cross-validation analysis. As predicted, the cross-validation errors were substantially lower for the Ensemble model than for the Fixed and Free IO models.

Can we conclude from these results that the decision variable endorsed by the Ensemble model better characterizes the decision-making process of an eyewitness? To address this question, we conducted model recovery analyses, using data from Wooten et al. (2020). Model recovery involves generating synthetic data from the evaluated models and then, fitting—in this case, cross-validating—each model to data generated from itself and from competing models (Myung & Pitt, 1997). Each model is expected to fit best to data generated from itself (i.e., the true model). If a given model can fit to data generated from another

**Figure 12**

*Empirical ROCs and the Best-Fitting ROC Curves for Each Variant of the Independent Observations Model for Experiments 1 and 2*



*Note.* Points denote the empirical data, and curves denote the model prediction based on best fitting values of  $\mu$  and  $\sigma$  for each model. Data and model fits in the showup procedure and the two-, four-, six-, and eight-person lineup procedures in Experiment 1 are denoted by black, dark gray, light gray, ash, and abalone shades, respectively. Data and model fits in the showup procedure and the three- and six-person lineup procedures in Experiment 2 are denoted by black, ash, and abalone shades, respectively. The graphs represent only the truncated ROC curves, that is, the proportion of guilty suspect identifications on target present trials and filler identification on target absent trials corrected by lineup size (Experiment 1) or innocent suspect identifications on target absent trials (Experiment 2). ROC = receiver-operating characteristic; FA = false alarm.

model better than the true model, that would attest to the undue flexibility of that model, which would limit the interpretability of the results from model fitting. The model recovery was conducted by simulating data for three-, six-, nine-, and 12-person lineups in which the strengths of the memory signals elicited by the guilty suspect and fillers were either uncorrelated or correlated (with values of 0.4 or 0.8). The decision rule of each model was applied to these data to compute the rates of each response type (target ID, innocent ID, no ID). Then, each model was cross-validated on the data simulated from each of the three models. The details of the methodology are presented in more detail in Appendix B. When the signals were uncorrelated, each model generalized best to the data generated from itself. However, when the correlation was set to 0.4 or 0.8, the Ensemble model provided the best fits to data generated from

not only itself but also from all other models. Table 7 shows mean error values ( $MSE_{CV}$ ) for each cross-validation analysis.

It is important to note that the incorporation of correlations between the targets and fillers represents a theoretical departure from the IO model, which assumes that the memory signals are independent of each other. We call the models that use the same decision variable as the IO models but with correlated data, the DO models. The *Fixed* DO model generates a lineup size pattern that is very similar to the pattern generated by the Ensemble model with correlated signals (see Figure 13). In other words, it is the correlations among the signals, and not the decision variable, that leads to an effect of lineup size on discriminability. Consequently, both the DO model and the Ensemble model provide an explanation for the overall pattern of results we (and Wooten et al., 2020) report.

The fact that the Fixed and Free IO models did not generalize to the data generated by the DO models as well as Ensemble model is instructive. After all, the DO models assume the same decision variable as the IO models, and likelihood functions of neither the Ensemble nor the IO models incorporate the correlations in the data. It thus appears that the Ensemble model can generate a lineup size effect that mimics a pattern generated by correlated data with a different decision rule.

However, there is no way of knowing whether the superiority of the Ensemble model is driven by the correlations in the data, rather than by its decision variable. A direct comparison of the Ensemble model with the DO model is needed to evaluate which decision rule better characterizes the eyewitness's decision process. Both the DO and the Ensemble models are better contenders than the IO model in characterizing lineup size effects. The comparison of these two models necessitates the derivation of likelihood functions that incorporate the correlations between targets and fillers. Unfortunately, this step complicates the mathematics and fitting algorithm considerably (Nadarajah et al., 2019), and is beyond the scope of what we can present here.

Overall, lessons from the model evaluations are threefold: (a) The Hybrid IO model received the most support, predicting the data better than the Fixed and Free IO models in Experiment 1, and performing as well or better than any other model in Experiment 2. This result confirms the empirical observation that the showup procedure leads to less diagnostic decisions compared with the lineup procedure even under particular assumptions about the distributions of underlying evidence. (b) For reasons elaborated at length, the Ensemble model is expected to be outperformed by the IO models when jointly modeling the showup and lineups of different sizes, and to outperform the IO model when modeling only the lineups. This was confirmed by the modeling of data from Experiments 1 and 2 and data from Wooten et al. (2020). (c) Comparing the IO model with the Ensemble model cannot reveal which decision variable better reflects the decision-making process of the eyewitness. This is because the effects of lineup size result from unmeasured correlations among the evidence values for the faces within a lineup. The varying decision variables of the two models produce similar data patterns when target and filler strengths are assumed to be correlated.

## General Discussion

In two experiments, we compared identification accuracy across the showup procedure and lineup procedures of varying sizes. We

**Table 3**

*Summary of Results of the Cross-Validation Analysis, as Well as the Fit Statistics Based on Each Model's Fit to the Entire Set of Data*

Model	Fixed Ensemble	Fixed IO	Hybrid IO	Free IO
Experiment 1 $N = 4,401$				
# of parameters	57	57	59	65
$-\ln(L)$	8709.9	8705	8698	8688
AIC	17534	17525	17515	<b>17507</b>
BIC	17898	<b>17889</b>	17892	17922
$MSE_{CV}$	.222 (.065)	.185 (.054)	<b>.181 (.063)</b>	.204 (.066)
Experiment 2 $N = 6,300$				
# of parameters	11	11	13	15
$-\ln(L)$	7691.2	7688	7685	7680
AIC	15405	15398	15395	<b>15390</b>
BIC	15479	<b>15473</b>	15483	15491
$MSE_{CV}$	.015 (.005)	<b>.012 (.005)</b>	<b>.012 (.005)</b>	.013 (.006)

*Note.* IO = independent observations; AIC = Akaike information criterion; BIC = Bayesian information criterion;  $MSE_{CV}$  = mean squared error of cross-validation. Values in bold indicate which model is favored by each measure of model performance. We interpret AIC values in the standard way (Burnham & Anderson, 2004), meaning that if a competing model has AIC values that are within two points of the best performing model, the competing and best performing model perform comparably.

used a measure of diagnostic accuracy that disentangles the influence of response bias on performance to compare lineups of different sizes. Experiment 1 used a between-subjects design; each participant made a single identification as is typical in eyewitness memory research. Experiment 2 adopted a within-subjects design; each participant made multiple identifications from lineups of different sizes.

Both experiments converged on the finding that showups lead to lower discriminability between the guilty and the innocent suspect compared to lineups. This finding replicated prior work that compared the two procedures using ROC analysis (Gronlund et al., 2012; Mickes, 2015; Wetmore et al., 2017; Wooten et al., 2020), indicating that the showup is an inferior procedure for soliciting eyewitness evidence.

The comparison of discriminability across lineups of different sizes, on the other hand, did not reveal statistically significant differences. However, in both experiments, the ROCs for larger lineups lay above the ROCs for smaller lineups. It is possible that the current experiments were not sufficiently powered to detect small differences in discriminability between lineups of different

sizes. However, a similar pattern was also observed by Wooten et al. (2020), with an even larger sample size.

Model-based analysis revealed that the quality of the information that is elicited from the guilty suspect in a showup is lower than the quality of information elicited by the guilty suspect in a lineup. This is inconsistent with an orthodox version of the IO model, in which the strength value for each face is independent of one another, and the raw maximum memory signal is compared with  $c$  criterion. The relative success of the *Hybrid IO* model suggests that showups are subject to different governing parameters than lineups, but that lineups differing in size are not.

Unlike the Ensemble model, the *Hybrid IO* model does not follow from a particular theoretical position. It is merely descriptive, and does not provide a mechanism that explains why performance is lower in showups than in lineups. As mentioned previously, the *Hybrid IO* model was implemented based on findings from prior and current work showing a disadvantage for the showup, and no significant changes in discriminability with increasing lineup size. The Ensemble model, on the other hand, is a mathematical instantiation of the *diagnostic feature-detection* the-

**Table 4**

*Best-Fitting Parameter Estimates of  $\mu$  and  $\sigma$  for Each Variant of the Independent Observations Model*

Experiment	Lineup size	Fixed Ensemble		Fixed IO		Hybrid IO		Free IO	
		$\mu$	$\sigma$	$\mu$	$\sigma$	$\mu$	$\sigma$	$\mu$	$\sigma$
Experiment 1	Showup	0.89	1.14	0.9	1.04	0.88	0.75	0.88	0.75
	Two-person					0.89	1.09	0.68	1.17
	Four-person							0.98	0.95
	Six-person							1.16	0.97
	Eight-person							0.72	1.22
Experiment 2	Showup	0.81	2.98	0.91	1.57	0.86	1.33	0.86	1.33
	Three-person					0.93	1.63	0.99	1.4
	Six-person							0.79	1.94

*Note.* IO = independent observations.

**Table 5**

*Proportion of Times ( $P_{CV}$ ) in Which the Model Listed on the Left Was Favored Over the Model Listed on the Right (i.e., Yielded a Lower  $SSE_{CV}$ ) Across 100 Cross Validation Iterations*

Model comparison	Experiment 1	Experiment 2
Fixed IO versus Ensemble	.69	.70
Hybrid IO versus Ensemble	.72	.70
Free IO versus Ensemble	.58	.56
Fixed IO versus Free IO	.74	.55
Hybrid IO versus Fixed IO	.57	.49
Hybrid IO versus Free IO	.64	.59

Note. IO = independent observations.

ory (Wixted & Mickes, 2014). When combined with the assumption that showups are governed by an IO process, it provides an explanation for the reduced discriminability in showups.

Overall, the model evaluation, through model recovery analyses and model fitting, revealed that the comparison of the Ensemble and IO models was not very informative in revealing which model better characterized the data. This study revealed that two independent factors can determine the effect of lineup size on discriminability—the correlation between targets and fillers, and the decision variable. When the data are correlated, the lineup size pattern is observed regardless of the operating decision variable. When the data are uncorrelated, the decision variable endorsed by the Ensemble model can generate the lineup size pattern. Thus, it appears that the manipulation of lineup size does not provide a felicitous test bed for comparing the Ensemble and IO models. However, further comparison of the DO model with the Ensemble model in terms of their predictions on lineup size effects does have the potential to provide insight into the operating decision variable. This would be a worthwhile endeavor for future research.

The comparison of the IO subvariants, on the other hand, provided support for the *Hybrid* model, which assumed different levels of performance across the showup and lineup procedures and roughly the same level of performance across lineups of different sizes. On the face, this appears to be little more than a restatement of the results from the empirical ROCs, but this is not accurate. Empirical ROCs are tremendously useful because their interpretation does not rely on theoretical assumptions about the distribution of evidence. Here the procedural inferiority of showups was demonstrated both by the atheoretical pAUC measure and

**Table 6**

*Summary of Results of the Cross-Validation Analysis, as Well as the Fit Statistics Based on Each Model's Fit to the Entire Set of Lineup Data From Wooten et al. (2020)*

Model	Fixed Ensemble	Fixed IO	Free IO
$N = 8,228$			
# of parameters	18	18	24
$-\ln(L)$	14429	14384	<b>14371</b>
AIC	28893	28804	<b>28789</b>
BIC	29019	<b>28930</b>	28957
$MSE_{CV}$	<b>.136 (.026)</b>	.233 (.035)	.227 (.028)

Note. IO = independent observations. Values in bold indicate which model is favored by each measure of model performance.

**Table 7**

*Results From the Model Recovery Simulations*

Evaluated model	Data generating (true) model		
	Ensemble	Fixed IO	Free IO
$\rho = 0$			
Ensemble	<b>.095 (.018)</b>	.159 (.032)	.171 (.036)
Fixed IO	.163 (.027)	<b>.146 (.021)</b>	.162 (.027)
Free IO	.187 (.029)	.150 (.024)	<b>.152 (.024)</b>
$\rho = .4$			
Ensemble	<b>.051 (.015)</b>	<b>.071 (.029)</b>	<b>.084 (.059)</b>
Fixed IO	.100 (.022)	.145 (.028)	.167 (.031)
Free IO	.090 (.021)	.121 (.023)	.127 (.023)
$\rho = .8$			
Ensemble	<b>.021 (.011)</b>	<b>.076 (.031)</b>	<b>.084 (.026)</b>
Fixed IO	.022 (.010)	.167 (.033)	.200 (.040)
Free IO	.024 (.011)	.149 (.030)	.150 (.030)

Note. IO = independent observations. Data were generated from the Ensemble, and Fixed and Free IO/DO models. Then, each model was fit to half of the data set simulated from each of the three models and cross-validated to the other half. The mean and standard deviation of the error values yielded by each cross-validation iteration are reported. Values in bold indicate the lowest error values ( $MSE_{CV}$ ).

also by cross-validation of models, which make theoretical assumptions about the distributions of underlying evidence and the decision-making process. This is important because empirical and theoretical discriminability may not always agree (Lampinen, 2016; Wixted & Mickes, 2018).

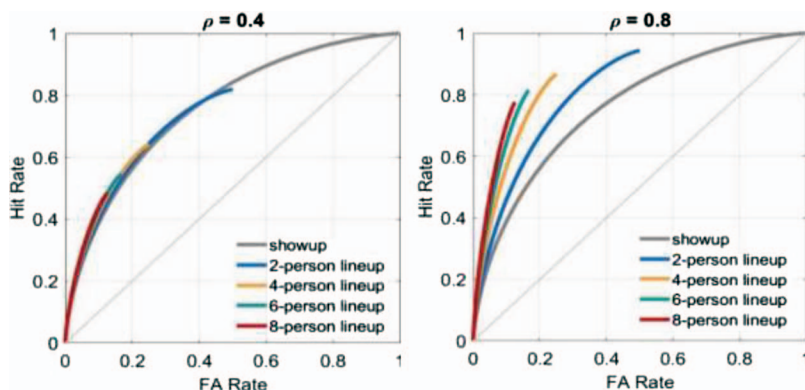
Though an explanation for the poorer performance in showups remains a target for further study, there are reasons to expect that result. According to the diagnostic feature-detection theory (Wixted & Mickes, 2014), lineups reveal to the eyewitness what the nondiagnostic and the diagnostic features are in the discrimination they face. For example, if all members of the lineup consist of tall, dark-haired males in their early 30s, the eyewitness will know not to base their identification decision on these features but to focus on features that differentiate the lineup members from each other. The Ensemble model is a mathematical implementation of that theory, and this idea is instantiated by its subtractive decision rule—the subtraction leads to the delineation of distinctive and potentially diagnostic features of each face. As such, the diagnostic feature-detection theory suggests that a showup may elicit lower quality information because it does not constrain the search of memory to maximally differentiating characteristics. Because a similar outcome obtains even with a decision rule based on raw memory strength signals, as embodied in the DO model, it is not currently possible to adjudicate which decision variable is a superior account of eyewitness decision-making.

The inclusion of a single filler in an identification procedure can instantly reveal many of the nondiagnostic features. Following our example above, the inclusion of a single description-matched filler would be sufficient in preventing the eyewitness from basing their identification decision on age, height, and hair color. An increase in lineup size would only be expected to increase discriminability to the extent that each additional filler provides new information that helps the identifier further constrain their memory search. However, as the lineup size increases, additional fillers might not



**Figure 13**

*Simulated ROC Curves for the Dependent Observations Model for the Showup, Two-, Four-, Six-, and Eight-person Lineup Procedures When Discriminability Was Equated Across the Procedures ( $d' = 1$ )*



*Note.* For the simulations with correlated signals ( $\rho = 0.4$  and  $\rho = 0.8$ ), the strength values of the innocent suspect and fillers were determined by the guilty suspect's memory strength. ROC = receiver-operating characteristic; FA = false alarm. See the online article for the color version of this figure.

contribute unique information. Even when additional fillers do provide additional useful information, keeping track of those constraints might tax the working memory of the eyewitness and be used less, or less efficiently.

More work is needed to directly compare theories that incorporate a role for familiarity across a lineup in the decision from ones that evaluate each face independently. It is evident from the present work that simultaneous lineups are often to be favored over showups as a forensic procedure, but that there is little evidence for an "ideal" lineup size. In some situations, it may be hard for detectives to find fillers that match a suspect on critical dimensions—imagine if the suspect is extremely tall or has very distinctive features—and the search for a prespecified number of fillers to construct a lineup may only delay the administration of the lineup or lower the lineup quality. This research suggests that little is to be gained by delaying administration beyond the time needed to gather high-quality fillers that are immediately available.

In actual forensic investigations, law enforcement and triers of fact should take into consideration not only the discriminability engendered by a procedure, but also the response bias or demand characteristics associated with the identification procedures they use. An investigator who uses a technique known to elicit liberal responding should recognize that the probability of a correct identification and a false identification are both higher. Response bias is relatively easy to modulate via instructions to an eyewitness or by titration of reported confidence, and this can be done in accordance with an investigator's goals and subjective costs of errors (see Gronlund & Benjamin, 2018).

We know from numerous lab and field studies that the showup procedure leads to a more liberal response bias (Behrman & Davey, 2001; Steblay et al., 2003) in addition to lower discriminability. It is thought that this occurs because the showup procedure is inherently suggestive and so creates a higher social pressure on the eyewitness to make an identification response. This demand characteristic was found to be intensified when the eye-

witnesses were led to believe that the showup is conducted as part of an actual criminal investigation, as opposed to a lab simulation (Eisen et al., 2017). In lab studies like the ones presented here, we cannot simulate the pressure of the demand characteristics that arise in actual eyewitness interviews (Dysart et al., 2006). This fact should be kept in mind when considering how to apply these results to an actual forensic investigation. It is certainly possible that lineups with a small number of fillers also lead to a liberal response bias, and the extent to which they do in actual forensic investigations has not been fully captured in the current set of experiments. Thus, small lineups should be conducted with the awareness that they may create demand characteristics similar to showups.

Other aspects of this work also limit its generalizability to forensic applications. Laboratory studies like this one provide a high level of homogeneity in viewing conditions and thus restrict the range of encoding characteristics that would be experienced across a population of eyewitnesses to an actual crime. Faces presented in the lineups were always identical to those that were viewed during study, a circumstance that never occurs in real life, where faces (and bodies) often change position, clothes, lighting, and hairstyle. The results presented here suggest that other than showups, lineups of different sizes elicit approximately equivalent discriminability, but that conclusion needs to be reassessed in procedures that relax the experimental control we have applied here and trade it off for ecological validity.

## References

- Albrecht, A. R., & Scholl, B. J. (2010). Perceptually averaging in a continuous visual world: Extracting statistical summary representations over time. *Psychological Science*, 21(4), 560–567. <https://doi.org/10.1177/0956797610363543>
- Beck, J. R. (1991). Decision-making studies in patient management: Twenty years later. *Medical Decision Making*, 11(2), 112–115. <https://doi.org/10.1177/0272989X9101100207>

- Behrman, B. W., & Davey, S. L. (2001). Eyewitness identification in actual criminal cases: An archival analysis. *Law and Human Behavior*, 25(5), 475–491. <https://doi.org/10.1023/A:1012840831846>
- Brewer, N., Caon, A., Todd, C., & Weber, N. (2006). Eyewitness identification accuracy and response latency. *Law and Human Behavior*, 30(1), 31–50. <https://doi.org/10.1007/s10979-006-9002-7>
- Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research*, 33(2), 261–304. <https://doi.org/10.1177/0049124104268644>
- Carlson, C. A., Gronlund, S. D., & Clark, S. E. (2008). Lineup composition, suspect position, and the sequential lineup advantage. *Journal of Experimental Psychology: Applied*, 14(2), 118–128. <https://doi.org/10.1037/1076-898X.14.2.118>
- Clark, S. E. (2003). A memory and decision model for eyewitness identification. *Applied Cognitive Psychology*, 17(6), 629–654.
- Clark, S. (2012). Costs and benefits of eyewitness identification reform: Psychological science and public policy. *Perspectives on Psychological Science*, 7(3), 238–259. <https://doi.org/10.1177/1745691612439584>
- Clark, S. E., Benjamin, A. S., Wixted, J. T., Mickes, L., & Gronlund, S. (2015). Eyewitness identification and the accuracy of the criminal justice system. *Policy Insights from the Behavioral and Brain Sciences*, 2(1), 175–186. <https://doi.org/10.1177/2372732215602267>
- Clark, S. E., Erickson, M. A., & Breneman, J. (2011). Probative value of absolute and relative judgments in eyewitness identification. *Law and Human Behavior*, 35(5), 364–380. <https://doi.org/10.1007/s10979-010-9245-1>
- Colloff, M. F., Wade, K. A., & Strange, D. (2016). Unfair lineups make witnesses more likely to confuse innocent and guilty suspects. *Psychological Science*, 27(9), 1227–1239. <https://doi.org/10.1177/0956797616655789>
- Duncan, M. (2006). *A signal detection model of compound decision tasks*. (Tech Note DRDC TR 2006–256). Defence Research and Development Canada.
- Dysart, J. E., Lindsay, R. C., & Dupuis, P. R. (2006). Show-ups: The critical issue of clothing bias. *Applied Cognitive Psychology*, 20(8), 1009–1023. <https://doi.org/10.1002/acp.1241>
- Eckstein, M. P., Thomas, J. P., Palmer, J., & Shimozaki, S. S. (2000). A signal detection model predicts the effects of set size on visual search accuracy for feature, conjunction, triple conjunction, and disjunction displays. *Perception & Psychophysics*, 62(3), 425–451. <https://doi.org/10.3758/BF03212096>
- Egan, J. P. (1958). *Recognition memory and the operating characteristic*. USAF Operational Applications Laboratory Technical Note.
- Eisen, M. L., Smith, A. M., Olaguez, A. P., & Skerritt-Perta, A. S. (2017). An examination of showups conducted by law enforcement using a field-simulation paradigm. *Psychology, Public Policy, and Law*, 23(1), 1–22. <https://doi.org/10.1037/law0000115>
- Glanzer, M., & Adams, J. K. (1985). The mirror effect in recognition memory. *Memory & Cognition*, 13(1), 8–20. <https://doi.org/10.3758/BF03198438>
- Gronlund, S. D., & Benjamin, A. S. (2018). The new science of eyewitness memory. *Psychology of Learning and Motivation*, 69, 241–284. <https://doi.org/10.1016/bs.plm.2018.09.006>
- Gronlund, S. D., Carlson, C. A., Neuschatz, J. S., Goodsell, C. A., Wetmore, S., Wooten, A., & Graham, M. (2012). Showups versus lineups: An evaluation using ROC analysis. *Journal of Applied Research in Memory & Cognition*, 1(4), 221–228. <https://doi.org/10.1016/j.jarmac.2012.09.003>
- Innocence Project. (2018). *Eyewitness misidentification*. Retrieved January 10, 2020, from <https://www.innocenceproject.org/dna-exonerations-in-the-united-states/>
- Kintsch, W. (1968). Recognition and free recall of organized lists. *Journal of Experimental Psychology*, 78(1), 481–487.
- Koen, J. D., Barrett, F. S., Harlow, I. M., & Yonelinas, A. P. (2016). The ROC Toolbox: A toolbox for analyzing receiver-operating characteristics derived from confidence ratings. *Behavior Research Methods*, 49(4), 1399–1406. <https://doi.org/10.3758/s13428-016-0796-z>
- Lampinen, J. M. (2016). ROC analyses in eyewitness identification research. *Journal of Applied Research in Memory & Cognition*, 5(1), 21–33. <https://doi.org/10.1016/j.jarmac.2015.08.006>
- Levi, A. M. (2007). Research note: Evidence for moving to an 84-person photo lineup. *Journal of Experimental Criminology*, 3(4), 377–391. <https://doi.org/10.1007/s11292-007-9042-0>
- Levi, A. M. (2012). Much better than the sequential lineup: A 120-person lineup. *Psychology, Crime & Law*, 18(7), 631–640. <https://doi.org/10.1080/1068316X.2010.526120>
- Levi, A. M., & Lindsay, R. C. L. (2001). Lineup and photo spread procedures: Issues concerning policy recommendations. *Psychology, Public Policy, and Law*, 7(4), 776–790. <https://doi.org/10.1037/1076-8971.7.4.776>
- Lusted, L. B. (1971). Decision-making studies in patient management. *The New England Journal of Medicine*, 284(8), 416–424. <https://doi.org/10.1056/NEJM197102252840805>
- Macmillan, N., & Creelman, C. D. (1991). *Detection theory: A user's guide*. Cambridge University Press.
- Macmillan, N., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). Erlbaum.
- Mansour, J. K., Beaudry, J. L., & Lindsay, R. C. L. (2017). Are multiple-trial experiments appropriate for eyewitness identification studies? Accuracy, choosing, and confidence across trials. *Behavior Research Methods*, 49(6), 2235–2254. <https://doi.org/10.3758/s13428-017-0855-0>
- Mickes, L. (2015). Receiver operating characteristic analysis and confidence-accuracy characteristic analysis in investigations of system variables and estimator variables that affect eyewitness memory. *Journal of Applied Research in Memory & Cognition*, 4(2), 93–102. <https://doi.org/10.1016/j.jarmac.2015.01.003>
- Mickes, L., Flowe, H. D., & Wixted, J. T. (2012). Receiver operating characteristic analysis of eyewitness memory: Comparing the diagnostic accuracy of simultaneous versus sequential lineups. *Journal of Experimental Psychology: Applied*, 18(4), 361–376. <https://doi.org/10.1037/a0030609>
- Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review*, 4(1), 79–95. <https://doi.org/10.3758/BF03210778>
- Nadarajah, S., Afuecheta, E., & Chan, S. (2019). On the distribution of maximum of multivariate normal random vectors. *Communications in Statistics Theory and Methods*, 48(10), 2425–2445. <https://doi.org/10.1080/03610926.2018.1465088>
- National Research Council. (2015). *Identifying the culprit: Assessing eyewitness identification*. National Academies Press.
- Neumann, M. F., Schweinberger, S. R., & Burton, A. M. (2013). Viewers extract mean and individual identity from sets of famous faces. *Cognition*, 128(1), 56–63. <https://doi.org/10.1016/j.cognition.2013.03.006>
- Nosworthy, G. J., & Lindsay, R. C. (1990). Does nominal lineup size matter? *Journal of Applied Psychology*, 75(3), 358–361. <https://doi.org/10.1037/0021-9010.75.3.358>
- Police and Criminal Evidence Act. (1984). *Codes of practice, Code D*. Retrieved from [https://www.gov.uk/government/uploads/system/uploads/attachment\\_data/file/253831/pace-code-d-2011.pdf](https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/253831/pace-code-d-2011.pdf)
- Police Executive Research Forum. (2013). *A national survey of eyewitness identification procedures in law enforcement agencies*. Retrieved from <https://www.policeforum.org/free-online-documents>
- Postman, L. (1950). Choice behavior and the process of recognition. *The American Journal of Psychology*, 63(4), 576–583.
- Pozzulo, J. D., Dempsey, J. L., & Wells, K. (2010). Does lineup size matter with child witnesses. *Journal of Police and Criminal Psychology*, 25(1), 22–26. <https://doi.org/10.1007/s11896-009-9055-x>

- Ratcliff, R., Sheu, C. F., & Gronlund, S. D. (1992). Testing global memory models using ROC curves. *Psychological Review*, 99(3), 518–535. <https://doi.org/10.1037/0033-295X.99.3.518>
- Report on the Prevention of Miscarriages of Justice. (2005). *Department of Justice: Federal/Provincial/Territorial Heads of Prosecutions Committee Working Group*. Department of Justice Canada. Retrieved from <https://www.justice.gc.ca/eng/rp-pr/cj-jp/ccr-rc/pmj-pej/pmj-pej.pdf>
- Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J. C., & Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, 12(1), 1–8.
- Robinson, M. M., Benjamin, A. S., & Irwin, D. E. (2020). Is there a K in capacity? Evaluating the discrete-slot model of visual short-term memory capacity. *Cognitive Psychology*. Advance online publication. <https://doi.org/10.1016/j.cogpsych.2020.101305>
- Rotello, C. M., & Chen, T. (2016). ROC curve analyses of eyewitness identification decisions: An analysis of the recent debate. *Cognitive Research: Principles and Implications*, 1(1), 1–10. <https://doi.org/10.1186/s41235-016-0006-7>
- Rotello, C. M., Heit, E., & Dubé, C. (2015). When more data steer us wrong: Replications with the wrong dependent measure perpetuate erroneous conclusions. *Psychonomic Bulletin & Review*, 22(4), 944–954. <https://doi.org/10.3758/s13423-014-0759-2>
- Smith, A. M., Wells, G. L., Lindsay, R. C. L., & Penrod, S. D. (2017). Fair lineups are better than biased lineups and showups, but not because they increase underlying discriminability. *Law and Human Behavior*, 41(2), 127–145. <https://doi.org/10.1037/lhb0000219>
- Stebay, N., Dysart, J., Fulero, S., & Lindsay, R. C. (2001). Eyewitness accuracy rates in sequential and simultaneous lineup presentations: A meta-analytic comparison. *Law and Human Behavior*, 25(5), 459–473. <https://doi.org/10.1023/A:1012888715007>
- Stebay, N., Dysart, J., Fulero, S., & Lindsay, R. C. L. (2003). Eyewitness accuracy rates in police showup and lineup presentations: A meta-analytic comparison. *Law and Human Behavior*, 27(5), 523–540. <https://doi.org/10.1023/A:1025438223608>
- Stovall v. Denno, 388 U.S. 293 (1967).
- Swets, J. A. (1986a). Form of empirical ROCs in discrimination and diagnostic tasks: Implications for theory and measurement of performance. *Psychological Bulletin*, 99(2), 181–198. <https://doi.org/10.1037/0033-2909.99.2.181>
- Swets, J. A. (1986b). Indices of discrimination or diagnostic accuracy: Their ROCs and implied models. *Psychological Bulletin*, 99(1), 100–117. <https://doi.org/10.1037/0033-2909.99.1.100>
- Wells, G. L. (1993). What do we know about eyewitness identification? *American Psychologist*, 48(5), 553–571. <https://doi.org/10.1037/0003-066X.48.5.553>
- Wells, G. L., & Lindsay, R. C. (1980). On estimating the diagnosticity of eyewitness nonidentifications. *Psychological Bulletin*, 88(3), 776–784. <https://doi.org/10.1037/0033-2909.88.3.776>
- Wells, G. L., & Olson, E. A. (2002). Eyewitness identification: Information gain from incriminating and exonerating behaviors. *Journal of Experimental Psychology: Applied*, 8(3), 155–167. <https://doi.org/10.1037/1076-898X.8.3.155>
- Wells, G. L., Rydell, S. M., & Seelau, E. P. (1993). The selection of distractors for eyewitness lineups. *Journal of Applied Psychology*, 78(5), 835–844. <https://doi.org/10.1037/0021-9010.78.5.835>
- Wells, G. L., Smalarz, L., & Smith, A. M. (2015). ROC analysis of lineups does not measure underlying discriminability and has limited value. *Journal of Applied Research in Memory & Cognition*, 4(4), 313–317. <https://doi.org/10.1016/j.jarmac.2015.08.008>
- Wetmore, S. A., McAdoo, R. M., Gronlund, S. D., & Neuschatz, J. S. (2017). The impact of fillers on lineup performance. *Cognitive Research: Principles and Implications*, 2(48), 1–13. <https://doi.org/10.1186/s41235-017-0084-1>
- Wickens, T. D. (2002). *Elementary signal detection theory*. Oxford University Press.
- Wixted, J. T. (2020). The forgotten history of signal detection theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 46(2), 201–233.
- Wixted, J. T., & Mickes, L. (2012). The field of eyewitness memory should abandon probative value and embrace receiver operating characteristic analysis. *Perspectives on Psychological Science*, 7(3), 275–278. <https://doi.org/10.1177/1745691612442906>
- Wixted, J. T., & Mickes, L. (2014). A signal-detection-based diagnostic-feature-detection model of eyewitness identification. *Psychological Review*, 121(2), 262–276. <https://doi.org/10.1037/a0035940>
- Wixted, J. T., & Mickes, L. (2015). ROC analysis measures objective discriminability for any eyewitness identification procedure. *Journal of Applied Research in Memory & Cognition*, 4(4), 329–334. <https://doi.org/10.1016/j.jarmac.2015.08.007>
- Wixted, J. T., & Mickes, L. (2018). Theoretical vs. empirical discriminability: The application of ROC methods to eyewitness identification. *Cognitive Research: Principles and Implications*, 3(1), 9–30. <https://doi.org/10.1186/s41235-018-0093-8>
- 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>
- Wixted, J. T., & Wells, G. L. (2017). The relationship between eyewitness confidence and identification accuracy: A new synthesis. *Psychological Science in the Public Interest*, 18(1), 10–65. <https://doi.org/10.1177/1529100616686966>
- Wooten, A. R., Carlson, C. A., Lockamy, 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*. Advance online publication. <https://doi.org/10.1002/acp.3644>
- Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, 12(6), 1100–1122. <https://doi.org/10.1177/1745691617693393>

(Appendices follow)

## Appendix A

### Confidence-Accuracy Characteristic (CAC) Analyses

#### Experiment 1

CAC analysis is used to assess the relationship between confidence in identification accuracy and objective accuracy. More specifically, within the context of lineup identifications, it plots the probability that a suspect identification is accurate as a function of confidence in the identification decision. It has been found that high-confidence identifications are highly accurate (Wixted & Wells, 2017). We examined this relationship here to see if differences across lineup sizes were apparent, as was reported by Mickes (2015). The confidence-accuracy functions are shown in Figure A1. The 11 levels of confidence (0, 10, 20 . . . , 100) used in Experiment 1 were binned into five levels as indicated in the figure. It can easily be seen that confident responses (90–100) are highly diagnostic—the vast majority of identifications are of the truly guilty suspect. These high-confidence accuracy values were compared across lineup sizes using bootstrapped estimates for variance. Specifically, we examined the distribution of difference scores, computed using 10,000 bootstrapped accuracy estimates for responses made with the highest level of confidence (90% and above) between each lineup procedure. Diagnostic accuracy at the highest level of confidence was lower for the showup than the lineup procedures; however, the differences did not reach statistical significance. The  $p$  values corresponding to a bootstrapped sample of differences between the showup and the

lineups of different sizes were 0.12, 0.12, 0.03, and 0.08 for two-, four-, six-, and eight-person lineup procedures, respectively. Taken together, this vector of low  $p$  values is suggestive that the confidence-accuracy relationship may be inferior for showups than for lineups, but that conclusion remains provisional.

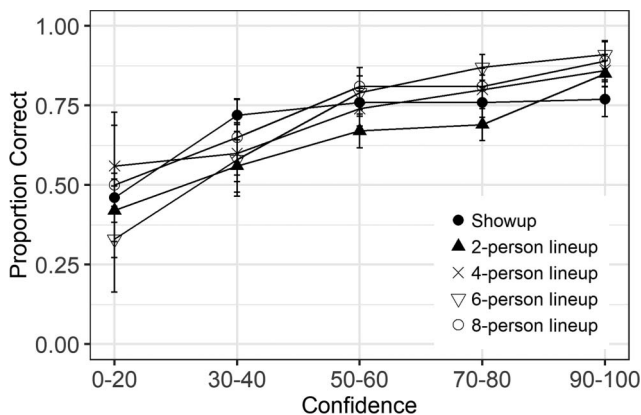
It is worth noting here that despite the large sample size adopted in this study, the calibration analyses were underpowered given the fewer number of data points that fell into each confidence bin. As previously explained, both the hit rates and the false rates decrease with increasing lineup size. Thus, the proportion correct scores are based on fewer data points with increasing lineup size (see Table 1).

#### Experiment 2

We also computed the confidence-accuracy function (see Figure A2) for the within-subject design used in Experiment 2. Participants who did not make any correct or false identifications within each level of confidence were excluded from analyses involving that confidence level. For both the showup and the lineup procedures, identification accuracy increased with the reported level of confidence. The level of accuracy for high-confidence responses was not lower for the showup procedure than the lineup procedures. This null result counters the pattern we observed in Experiment 1

**Figure A1**

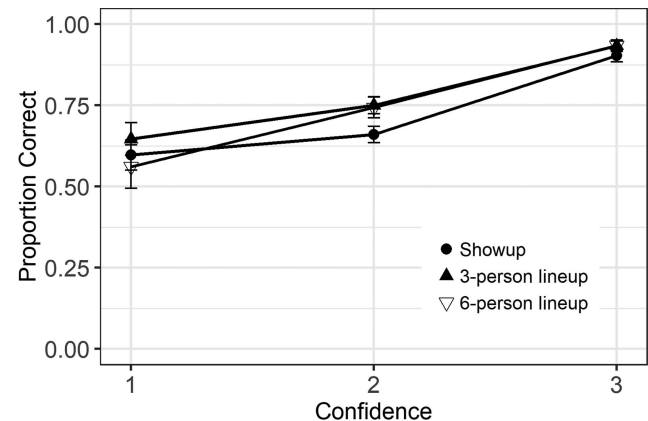
*Confidence-Accuracy Characteristic (CAC) Curves*



*Note.* Error bars represent standard error of the proportion correct values. The standard error was estimated using 10,000 bootstrap replicates.

**Figure A2**

*Confidence-Accuracy Characteristic (CAC) Curves for Showups, 3-person, and 6-person Lineups*



*Note.* Error bars represent standard error of proportion correct values. The standard error was estimated using 10,000 bootstrap replicates.

(Appendices continue)



and also the one reported by Mickes (2015) (Experiment 2) that high confidence responses are less accurate for showups than the lineups. Both of those prior studies had participants make a single decision. The differing pattern of results of the current experiment might be due to its within-subjects nature. Experiencing the lineup task might have helped participants to adjust an otherwise inflated confidence level for the showup procedure. However, further research is needed to evaluate this possibility.

## Summary

In both experiments, the accuracy of responses increased with increasing levels of confidence for both the showup and the lineup procedures and the responses made with the highest level of

confidence were diagnostic of accuracy. The probability of the identified suspect being guilty was very high for all lineup conditions, attesting to the utility of confidence ratings in identification decisions in the legal system. This pattern was less clearly observed for the showup procedure in Experiment 1, a result that replicates prior work (Mickes, 2015). However, that exception for showups was not replicated in Experiment 2. The high-confidence responses were as accurate for the showup procedure as for the lineup procedures. This difference might be due to procedural differences across the two experiments. Unlike in Experiment 1, participants made multiple responses in Experiment 2, an opportunity that may have allowed participants to more successfully calibrate their level of confidence to their accuracy.

## Appendix B

### Model Recovery Analysis

We used model recovery analysis to assess how our primary index of model comparison ( $MSE_{CV}$  from cross-validation) performed in recovering the true data-generating model (for other examples of model recovery analysis see Robinson et al., 2020). As explained in the main text, all model recovery was performed after excluding the showup procedure. We first obtained best fitting parameters by separately fitting the *Fixed* and *Free* variants of the IO model and the Ensemble model to the data from Wooten et al. (2020). We then used these best fitting parameters to simulate synthetic data from each of the three models. The number of observations in each lineup Size  $\times$  Target Presence condition was based on the sample sizes used by Wooten et al. (2020). We then evaluated each of the three models on each of the three data sets generated from these models using fivefold cross-validation anal-

ysis. The cross-validation analysis was performed by splitting the data into five folds, each of which had approximately an equal number of observations. Subsequently, each model was fit to four folds of the data generated from each of the three models—the training set, and best fitting parameters from these fits were used to predict the data in the fifth fold (i.e., the validation set). This procedure was repeated five times, using each fold as the validation set, and furthermore, the procedure was repeated 100 times, with 100 random splits of the data. The results from this analysis are reported in Table 6.

Received September 12, 2019

Revision received October 5, 2020

Accepted October 12, 2020 ■