Why are lineups better than showups? A test of the filler siphoning and enhanced discriminability accounts

Colloff, Melissa; Wixted, John T

DOI:
10.1037/xap0000218

Document Version
Peer reviewed version

Citation for published version (Harvard):
Colloff, M & Wixted, JT 2019, 'Why are lineups better than showups? A test of the filler siphoning and enhanced discriminability accounts', *Journal of Experimental Psychology: Applied*, vol. 26, no. 1, pp. 124-143. https://doi.org/10.1037/xap0000218

Why are Lineups Better than Showups? A Test of the Filler Siphoning and Enhanced Discriminability Accounts

Melissa F. Colloff & John T. Wixted

Author Note

Melissa F. Colloff, Centre for Applied Psychology, School of Psychology, University of Birmingham, UK; John T. Wixted, Department of Psychology, University of California, San Diego, CA.

Our data are freely available (https://osf.io/kyvqa/).

Sections of these data were presented by Melissa F. Colloff at the British Psychological Society Cognitive Section Conference (August, 2017), Newcastle University, UK and the 59th Annual Meeting of the Psychonomic Society (November, 2018), New Orleans, Louisiana, USA.

We thank Kimberley Wade and Sophie Nightingale for their helpful comments on an earlier draft. This work was supported by the National Science Foundation under Grant SES-1456571 (to J.T.W.) and by Postgraduate Study Visit Grants from the Experimental Psychology Society and the British Psychological Society.

Correspondence concerning this article should be addressed to Melissa F. Colloff, School of Psychology, University of Birmingham, Birmingham, UK, B15 2TT. Email: M.Colloff@bham.ac.uk

Abstract

Presenting the police suspect alongside similar-looking people (a lineup) results in more accurate eyewitness identification decisions than presenting the suspect alone (a showup). But why are lineups better than showups? *Diagnostic-feature-detection* theory suggests that lineups enhance witnesses' ability to discriminate between innocent and guilty suspects, because facial features can be compared across lineup members. *Filler-siphoning* suggests that the presence of other lineup members siphons some of the incorrect identifications that would otherwise land on the innocent suspect. To test these two accounts, over 3,600 subjects across three experiments watched a mock-crime video and were presented with either a showup, a simultaneous lineup, or a simultaneous showup (a novel procedure). Subjects in the simultaneous showup condition saw the suspect and five similar-looking faces, but, unlike a lineup, could not identify the other faces. Presenting similar-looking faces alongside the suspect (simultaneous showup and lineup) enhanced subjects' ability to discriminate between innocent and guilty suspects compared to presenting the suspect alone (showup) as measured by *p*AUC and fitting a signal-detection model. These results show, for the first time, that the discriminability advantage in simultaneous lineups is due to the comparison of multiple faces as predicted by diagnostic-feature-detection theory, but not the filler-siphoning account.

KEYWORDS: lineups, showups, diagnostic-feature-detection, filler siphoning, signal-detection theory

PUBLIC SIGNIFICANCE STATEMENT: Our research found that lineups result in more accurate eyewitness identifications than showups, because the opportunity to compare across similar-looking faces in a lineup boosts identification accuracy. Our work demonstrates how scientists and policy makers could use well-specified psychological theories to enhance existing, or develop new, eyewitness identification procedures to improve eyewitness accuracy.

Two identification procedures commonly used by the police are lineups and showups (Police Executive Research Forum, 2013). In a lineup, the police suspect is presented alongside other similar-looking individuals—fillers—who are known to be innocent. The lineup images are either presented one at a time (a sequential lineup) or all at once (a simultaneous lineup). In a showup, the police suspect is presented alone. In many countries, courts and legal scholars have criticized the use of showups, deeming them to result in unreliable eyewitness identifications (e.g., Stovall v. Denno, 1967; Wolchover & Heaton-Armstrong, 2014). This belief accords with the results of many empirical studies that have been interpreted to mean that lineups foster more accurate identifications than showups (Lindsay, Pozzulo, Craig, Lee, & Corber, 1997; Steblay, Dysart, Fulero, & Lindsay, 2003, but see Clark & Godfrey, 2009; Valentine, Davis, Memon, & Roberts, 2012). Recent studies using Receiver Operating Characteristic (ROC) analysis also support this impression because they have consistently found that *simultaneous* lineups yield better empirical discriminability—defined as the ability to tell the difference between innocent and guilty suspects—than showups (Neuschatz et al., 2016). Compared to showups, simultaneous lineups enhance empirical discriminability as measured by the partial Area Under the ROC Curve (*p*AUC), and this is true for both young and older adults (Gronlund et al., 2012; Key et al., 2015). Even simultaneous lineups that are delayed by 48 hours result in a larger *p*AUC (i.e., a higher *p*ROC) than showups conducted immediately (Wetmore et al., 2015). Note that there have been only two studies comparing *sequential* lineups to showups using ROC analysis. One used an unfair sequential lineup and found that the sequential lineup was superior to a showup (Gronlund et al., 2012). The other used a fair sequential lineup and found that the two procedures were comparable, with the showup exhibiting a slight but non-significant advantage (Wilson, Donnelly, Christenfield & Wixted, in press). Thus, empirical discriminability may not be better in sequential lineups than showups, because sequential lineup performance depends on a complex interplay of factors, such as the suspect's position in the lineup, the similarity of the fillers to the witness's memory of the culprit, and the witness's decision criterion (Wilson et al., in press). Because research comparing sequential lineups and showups using ROC analysis is scant and position effects in sequential lineups are complex, in this paper we focus on the well-established and well-accepted simultaneous lineup advantage over showups.

The fact that simultaneous lineups yield a larger *p*AUC than showups is not currently under dispute, but the theoretical interpretation of that effect is. Two interpretations have been offered: diagnostic-feature-detection theory (Wixted & Mickes, 2014) and filler siphoning (Wells, 2001; Wells, Smalarz, & Smith, 2015). Diagnostic-feature-detection theory was developed to account for the findings that diagnostic accuracy is higher for simultaneous lineups compared to sequential lineups and showups (both of which involve faces presented in isolation). Diagnostic-feature-

detection theory holds that simultaneous lineups enhance witnesses' underlying (i.e., theoretical) ability to discriminate innocent from guilty suspects compared to showups. According to this account some facial features differ between innocent and guilty suspects and are therefore diagnostic of guilt, whereas other facial features are shared by innocent and guilty suspects and are therefore non-diagnostic. The non-diagnostic features are those that correspond to the description of the perpetrator provided by the eyewitness and that are used to select fillers. Whether innocent or guilty, the suspect will have those features, which means that relying on those features to decide whether or not the perpetrator is in the lineup will harm performance. Critically, simultaneous lineups afford witnesses the opportunity to immediately appreciate which facial features are shared by all lineup members and are therefore non-diagnostic. In other words, in a fair lineup, the fillers and the suspect all match the description of the perpetrator, so presenting their photos simultaneously accentuates the non-diagnostic features. Witnesses are then able to discount the non-diagnostic features from their identification decision, which, in turn, enhances witnesses' ability to discriminate between innocent and guilty suspects. By contrast, showups do not permit comparison across multiple faces and therefore deny witnesses the opportunity to learn which facial features are shared. Witnesses may therefore rely to a greater extent on non-diagnostic features, which will impair their ability to discriminate between innocent and guilty suspects.

Another account holds that simultaneous lineups do not enhance witnesses' ability to discriminate between innocent and guilty suspects (Wells, 2001; Wells, Smalarz, & Smith, 2015). Instead, according to this account, the presence of plausible alternatives (fillers) in lineups siphons some of the incorrect identifications that would otherwise land on the innocent suspect. This *filler siphoning* phenomenon occurs to a greater extent in lineups that contain an innocent suspect (target-absent lineups) compared to lineups that contain a guilty suspect (target-present lineups) because, if the target-absent lineup is fair, no one stands out as generating a strong memory match signal. Thus, false identifications of the innocent suspect (being spread out across the fillers) are reduced to a greater extent than correct identifications of the guilty suspect, perhaps elevating the ROC. The theoretical mechanism that explains *why* the ROC should increase rather than remain unchanged has not been specified. Such a mechanism will be needed at some point because the filler siphoning phenomenon (false IDs being reduced to a proportionately greater extent than correct IDs) would be observed even if the ROC were not elevated (Colloff, Wade, Strange & Wixted, 2018). That caveat aside, this account holds that filler siphoning is protective of innocent suspects and in such a way that the ROC is elevated. Showups, however, do not provide this protection, simply because there are no fillers, so all of the identification errors that occur land on the innocent suspect. According to this interpretation, the larger *p*AUC for simultaneous lineups compared to showups occurs despite the

fact that the underlying theoretical ability to discriminate innocent from guilty suspects is the same for lineups and showups (see Wixted & Mickes, 2018, for a discussion of empirical vs. theoretical discriminability).

Both the diagnostic-feature-detection and filler siphoning accounts are consistent with the observation that fair simultaneous lineups in which all of the lineup members match the description of the suspect yield a higher *p*ROC than unfair simultaneous lineups in which the suspect stands out because he is more similar to the participant's memory of the real perpetrator than the other lineup members (Colloff, Wade, & Strange, 2016; Colloff, Wade, Wixted, & Maylor, 2017; Wetmore et al., 2015). Again, however, the two accounts differ in their explanation of why this effect occurs. According to the diagnostic-feature-detection theory, only a fair lineup affords the witness the opportunity to discern which features are shared and should not be used for the identification. An unfair lineup does not because in an unfair lineup the innocent suspect has features that do not match the fillers but do match the witness's memory of the guilty perpetrator. These features will not be discounted and will instead be interpreted by the witness as evidence that the innocent suspect is guilty. As such, the diagnostic-feature-detection theory predicts that the fair lineup produces a larger *p*AUC, because underlying theoretical discriminability—ability to discriminate between innocent and guilty suspects—is better in fair than unfair lineups. According to the filler siphoning account, in an unfair lineup, the suspect is more similar to the witness's memory of the perpetrator than the other fillers so the identifications are more likely to land on the suspect (whether innocent or guilty). As such, the filler siphoning account suggests that the fair lineup could yield a higher *p*AUC, even though there is no improvement in underlying theoretical discriminability in the fair lineup (see Smith, Wells, Smalarz, & Lampinen, 2018).

To be clear, the filler siphoning phenomenon does not *necessarily* predict that fair lineups yield a higher *p*AUC than showups. As shown by Colloff et al. (2018), introducing fillers will indeed siphon identifications away from innocent suspects to a greater extent than they siphon identifications away from guilty suspect (i.e., the false identification rate will decrease to a greater extent than the correct identification rate). However, the same phenomenon—a greater reduction in the false identification rate compared to the correct identification rate—occurs when responding becomes more conservative (e.g., Rotello & Chen, 2016; Rotello, Heit & Dubé, 2015; Wixted & Mickes, 2018). Thus, filler siphoning could disproportionately reduce correct and false identification rates without any change in *p*AUC. Put another way, filler siphoning (having the same effect as more conservative responding) could shift performance to a lower point on the same ROC curve, without moving it to a different, higher curve. Still, the argument has been made that filler siphoning does more than that and can also explain why *p*AUC is higher for fair lineups than showups. Throughout

this article, we use filler siphoning *theory* to refer to the idea that there is something about the filler siphoning phenomenon that does more than shift the correct and false identification rates to a more conservative position on the same ROC and instead elevates the ROC (thereby increasing *p*AUC).

Which theory—diagnostic-feature-detection or filler siphoning theory—best accounts for the higher *p*AUC observed for simultaneous lineups compared to showups? There is currently no empirical evidence addressing this question. Here, we pit the two accounts against each other. Both theories predict that performance will differ when the suspect is presented in a lineup compared to when the suspect is presented in a showup, but the diagnostic-feature-detection model suggests this effect occurs because the presence of other similar-looking faces enhances people's ability to discriminate between innocent and guilty suspects, while the filler siphoning account suggests this effect occurs because the other similar-looking faces attract some of the erroneous identifications. Critically, then, the diagnostic-feature-detection theory predicts that presenting similar-looking faces alongside the suspect (like a lineup) will enhance witnesses' discriminability (i.e., increase *p*AUC and *d'*) compared to presenting the suspect alone, even when there is no opportunity for filler identifications. The filler siphoning account predicts no benefit (i.e., no increase in *p*AUC) of presenting similar-looking faces alongside the suspect if there is no opportunity for filler identifications.

Experiment 1

To test these predictions, our subjects watched a mock crime video and were presented with either a *simultaneous showup* (a novel procedure) or a *standard showup*. Subjects in the simultaneous showup condition saw the suspect and five similar-looking faces, but, unlike a standard simultaneous lineup, were told that the other five similar-looking faces were not suspects and were prevented from identifying these other faces. Instead, in both conditions, the suspect was highlighted and subjects were asked whether this was the person who committed the crime. The diagnostic-feature-detection hypothesis predicts better discriminability—better ability to discriminate between innocent and guilty suspects—in the simultaneous showup than the standard showup, but the filler siphoning account does not. Note that the diagnostic-feature-detection hypothesis predicts better discriminability even in situations when the similar-looking faces are not considered to be possible suspects (i.e., in our simultaneous showup), because the extraction and discounting of common features is thought to be an automatic phenomenon. In the literature on ensemble coding, for example, the standard assumption is that summary statistics are quickly and automatically computed whenever a set of similar objects are simultaneous presented (Ariely, 2001; Whitney & Yamanashi-Leib, 2018). Thus, it is assumed that the subject would, without deliberate effort, appreciate the

degree to which the suspect 'stands out' from the surrounding faces. Basing a decision on that variable (i.e., the degree to which the suspect face 'stands out' from the surrounding faces) would theoretically enhance discriminability. This ensemble-coding version of the diagnostic feature-detection hypothesis was recently tested and supported in a study reported by Wixted, Vul, Mickes and Wilson (2018), and makes the predictions that we test in the three Experiments in this paper.

Method

Design

We used a 2 (presentation: simultaneous showup, standard showup) × 2 (video: mugging, graffiti) × 2 (target: present, absent) mixed design, with video and target manipulated within subjects. This mixed design enabled us to collect a second data point per subject. Methods for calculating a priori power analysis for eyewitness identification experiments are not well specified, but ROC lineup studies usually recruit between 300 and 500 subjects per condition. Our data-collection stopping rule was to recruit at least 1,000 subjects—500 in each of the between-subjects conditions. Using the mean difference and standard deviations observed in Wetmore et al. (2015) as a guide, a power analysis indicated that, with 500 subjects per between-subject condition, power for this showup experiment would exceed 80%. The research was reviewed according to the University of California, San Diego IRB procedures for research involving human subjects.

Subjects

The subjects were 1,130 undergraduates from the University of California, San Diego (UCSD) who received course credit for participating in the experiment. We excluded 89 people (7.88% in total) who had completed the experiment more than once ($n = 46$), experienced technical difficulties while watching the video ($n = 8$), or incorrectly answered an attention check question on the content of the video ($n = 35$). This resulted in a final sample of 1,041. Table 1 shows a demographic breakdown of the sample.

Table 1
Subject Demographic Information for Experiments 1, 2, and 3

	Experiment 1	Experiment 2	Experiment 3
Sex			
Male	262	355	558
Female	772	639	1,072
Other	0	0	3
Prefer not to say	7	9	9
Age			
M	20.27	20.28	19.88
SD	2.73	2.42	2.10
Prefer not to say	2	6	3
Ethnicity			
African-American	22	15	33
Asian	563	479	854
Caucasian	189	195	298
Filipino	38	42	52
Latino	46	52	102
Mexican-American	86	137	179
Native-American	0	0	2
Other	78	66	105
Prefer not to say	19	17	17

Materials

We used two 30 s videos depicting different non-violent crimes—a mugging and a graffiti attack. Each perpetrator had a distinctive facial feature, either a tattoo (mugging) or a black-eye (graffiti).

Showups

Colloff et al. (2016) compiled a pool of 40 fillers for each perpetrator. We randomly selected 12 of these fillers (6 for each perpetrator) for our study. The perpetrator's distinctive feature had been digitally added to each of the filler faces because this is one method of constructing fair lineups for distinctive suspects (Colloff et al., 2016; Zakardi, Wade, & Stewart, 2009). The stimuli have been piloted tested (see Colloff et al., 2016). Five subjects examined the stimuli and were satisfied that all of the final fillers matched the descriptions of the perpetrators and did not look like they had been digitally altered. Moreover, a new group of subjects ($N = 39$) viewed a 6-person target-present lineup for each perpetrator in which the fillers were randomly selected from the pool of fillers, and were asked to identify which photograph had not been digitally altered. The proportion of subjects who selected the perpetrator was not significantly different from chance (chance = 17%; graffiti: 17.9% picked the perpetrator, $t(38) = 0.206$, $p = .84$; mugging: 12.8% picked the perpetrator, $t(38) = 0.709$, $p = .483$). Together, this result indicates that people were unable to tell the difference between the

perpetrators who had distinctive features during filming and the fillers who had digitally added distinctive features.

For each perpetrator, we randomly selected one person to be the designated innocent suspect. Therefore, our simultaneous showups either consisted of the perpetrator and five fillers (target-present) or the innocent suspect and five fillers (target-absent), and our standard showups were either a single photo of the perpetrator (target-present), or the innocent suspect (target-absent). To check whether the innocent suspect and fillers in each of our simultaneous showups were plausible alternatives to the perpetrators, we conducted a standard mock-witness test and provided a group of mock-witnesses with a modal description of the perpetrator (created by subjects in the study by Colloff et al., 2016) and either a target-present or target-absent simultaneous *lineup* for that perpetrator. We refer to these as lineups, because the mock-witnesses were not aware of whom the suspect was and were allowed to pick any face. To be clear, the mock-witnesses did not view the mock crime videos; their task was simply to pick the person in the lineup that they deemed to best fit the description they had read. As such, mock-witness testing determines whether one or more lineup members are *perceptually* distinct from the other lineup members, based on a description of the perpetrator. Forty different mock-witnesses viewed each simultaneous lineup (total $N = 160$). We calculated Tredoux's E', which uses the distribution of mock-witness choices to determine how many lineups members are appropriate (i.e., it measures effective size; Tredoux, 1999). For the mugging scenario, Tredoux's E' was 4.17 (95% CI [3.47, 5.22]) for the target-present lineup, and was 3.79 (95% CI [3.22, 4.62]) for the target-absent lineup. For the graffiti scenario, Tredoux's E' was 4.37 (95% CI [3.53, 5.75]) for the target-present lineup, and was 3.92 (95% CI [3.12, 5.27]) for the target-absent lineup. This indicates that in each lineup there were approximately 4 members who were viable alternatives from which the witness might choose. In the mugging scenario the perpetrator and innocent suspect were chosen by 30% and 22.5% of the mock-witnesses, respectively. In the graffiti scenario the perpetrator and innocent suspect were chosen by 25% and 17.5% of the mock-witnesses, respectively. Taken together, these values were considered acceptable because they compare favourably against estimates from field studies. Valentine and Heaton (1999), for example, found that in a sample of 9-person photo and video lineups in the UK the average effective size ranged from 4.24 to 4.46 and the proportion of mock-witnesses selecting the suspect was between 12%-25% (chance was 11%). Our values are also comparable to other laboratory studies that concluded that their lineups were fit for purpose, such as Horry, Palmer, and Brewer (2012) who found across 11 6-person lineups the average effective size ranged from 3.69 to 3.75, and the proportion of mock-witnesses selecting the suspect was between 19% and 28% (chance was 17%).

Procedure

Subjects were told that the study was about perception and memory and were randomly assigned into conditions. First, subjects watched a video of a crime (either mugging or graffiti). They were instructed to pay close attention because they would be asked questions about it later. After the video had finished, we checked whether subjects had encountered any technical problems, such as excessive buffering. Next, subjects completed a filler task, in which they attempted to solve spatial reasoning questions for 4 min. Following this, subjects were asked to rate their confidence that they would be able to recognize the perpetrator from the video on an 11-point Likert-type scale ranging from 0 (*completely uncertain*) to 100 (*completely certain*). Next, all subjects were told that they would be asked to decide whether the police suspect was the male perpetrator from the video. Those in the simultaneous showup condition were told that they would be presented with a lineup of images, while those in the standard showup condition were told that they would be presented with a photo. All subjects were instructed to look at the photo(s) carefully, and, after 10 seconds, further instructions would appear.

The identification task was displayed on the next screen. In the simultaneous showup condition, six faces were displayed simultaneously in two rows of three photos. Subjects in the target-present condition saw the perpetrator and five fillers, while subjects in the target-absent condition saw the innocent suspect and five fillers. The order of the faces was randomly generated. After 10 s, a thick red border appeared around the suspect—the perpetrator in the target-present condition, or the innocent suspect in the target-absent condition—and additional instructions were displayed. Subjects were told: *"The police suspect is highlighted in red. The other five men are not suspects; their role is to help you decide whether the suspect is the person that committed the crime. The police suspect may or may not be the actual perpetrator."* In the standard showup condition, one photo was displayed. Subjects in the target-present condition saw the perpetrator, while subjects in the target-absent condition saw the innocent suspect. After 10 s, a thick red border appeared around the image and additional instructions were displayed. Subjects were told: "*This is the police suspect. The police suspect may or may not be the actual perpetrator*." All subjects were asked the same question: *"Is the suspect (highlighted in red) the person who committed the crime?"* and responded by clicking on "Yes" or "No". Following this, subjects used an 11-point Likert-type scale (0=*completely uncertain* to 100=*completely certain*) to rate their confidence in their decision and answered a question that enabled us to check that they were paying attention ("How many people were in the video?").

The procedure then began again, but this time subjects were allocated into the alternate video (mugging or graffiti) and target (present or absent) condition. The order of the video and target

conditions was counterbalanced. Finally, at the end of the study, subjects answered a number of demographic questions.

Results & Discussion

Our aim to was to determine whether presenting similar-looking faces alongside the suspect (a simultaneous showup) enhances witnesses' ability to discriminate between innocent and guilty suspects compared to presenting the suspect alone (a standard showup). We addressed this question using ROC analysis. Our data are freely available (https://osf.io/kyvqa/).

At this juncture, it is important to consider that eyewitness ROC studies comparing showups and simultaneous lineups have analyzed the performance of subjects who responded "Yes, that is the culprit" (i.e., choosers); they have not analyzed the performance of subjects who responded "No, the culprit is not here" (i.e., non-choosers). This is because ROC lineup studies to date have only required that choosers—and not non-choosers—rate their confidence that an *individual* (i.e., the person that they identified) is the perpetrator. Only partial ROC (*p*ROC) curves have been constructed and, as such, the theoretical debate about the possible filler-siphoning and diagnostic-feature-detection mechanisms has focused on the performance of choosers. Indeed, the diagnostic-feature-detection theory was developed to account for the findings of studies plotting *p*ROC curves (Wixted & Mickes, 2014). Nevertheless, ROC studies examining showups can also allow for the analysis of non-choosers by constructing full ROC curves. This is because showup studies can also ask non-choosers to rate their confidence that the *individual* presented (i.e., the suspect) is not the perpetrator. Although we had planned to analyze the Yes responses in accordance with the previous literature, improving theoretical understanding of non-choosers and how they might differ from choosers is also important. Therefore, for our showup experiments (Experiments 1 and 2), we report *p*ROC analyses for Yes responses (chooser data) and also full ROC analyses for Yes and No responses (chooser and non-chooser data).

ROC Analysis

*p*ROC curves have been plotted extensively in the lineup literature (see Gronlund, Wixted, & Mickes, 2014; Mickes, Flowe, & Wixted, 2012). To construct our *p*ROC curves, we took our 11-point confidence scale, ranging from Yes 100 to Yes 0, and plotted the cumulative hit rate (HR; number of Yes responses to the guilty suspect ÷ number of target-present lineups) against the cumulative false alarm rate (FAR; number of Yes responses to the innocent suspect ÷ number of target-absent lineups) over decreasing levels of confidence. Looking at Figure 1A, the leftmost point on the *p*ROC includes only those chooser IDs made with the highest level of confidence (Yes 100). The next point includes only those chooser IDs made with highest and the second-highest level of confidence (Yes 100 and Yes 90). The rightmost point on the *p*ROC includes chooser IDs made with

any level of confidence (Yes ≥ 0). To construct our full ROCs, we formed a single 21-point confidence scale, ranging from Yes 100 to Yes 0 then No 0 to No 100. We collapsed Yes 0 and No 0 into one category. The full ROC plots the cumulative hit rate (HR; number responses to the guilty suspect ÷ number of target-present lineups) against the cumulative false alarm rate (FAR; number responses to the innocent suspect ÷ number of target-absent lineups) over decreasing levels of confidence. As such, the left side of the full ROCs in Figure 1B match exactly the *p*ROCs in Figure 1A. The full ROCs in Figure 1B simply extend the *p*ROCs to also take into account gradations of confidence in No responses (i.e., the non-chooser data).

In both partial and full ROC analysis, the procedure with the ROC curve that falls furthest from the dashed chance line is best at enhancing empirical discriminability—people's collective ability to discriminate between guilty and suspect. This is because the procedure with the higher ROC results in more guilty suspect IDs and fewer innocent suspect IDs than the alternative procedure. *p*ROC curves are compared statistically by computing the partial Area Under the Curve (*p*AUC). Full ROC curves are compared statistically by computing the Area Under the Curve (AUC). We used the statistical package *pROC* to calculate (*p*)AUC and *D,* a measure of effect size: $D = (AUC1 - AUC2)/s$, where *s* is the standard error of the difference between the two AUCs and is estimated using bootstrapping (Robin et al., 2011). In all *p*AUC analyses, we defined the specificity (1 – FAR) using the smallest false alarm rate (FAR) range in that comparison.

Collapsed over both videos. Figure 1A shows the *p*ROC curves for the simultaneous showup and standard showup, collapsed over the two mock crime videos. It is clear from Figure 1A that the *p*ROC curves lie directly on top of each other. This indicates that, in the aggregate, we did not find the simultaneous showup advantage predicted by the diagnostic-feature-detection theory. Indeed, the *p*AUC (specificity = .68) for the simultaneous showup (*p*AUC = .12, 95% CI [.11, .14]) was nearly identical to, and so was not significantly greater than, the *p*AUC for the standard showup, *p*AUC = .12, 95% CI [.11, .14], $D = 0.14$, $p > .250$. Figure 1B shows that the same pattern is observed when the full ROCs are plotted. The AUC for the simultaneous showup (AUC = 0.70, 95% CI [0.67, 0.73]) was nearly identical to the AUC for the standard showup, AUC = 0.71, 95% CI [0.68, 0.75], $D = 0.73$, $p > .250$.

Separated by video. We next analyzed identification performance separately for each video. It was immediately apparent that identification accuracy for the mugging video (simultaneous showup $d' = 0.25$, standard showup $d' = 0.59$) was much poorer than accuracy for the graffiti video (simultaneous showup $d' = 1.48$, standard showup $d' = 1.12$; for a discussion of the relationship between *p*AUC and *d′,* see Mickes, Moreland, Clark, & Wixted, 2014). Figure 1C-F show the partial and full ROC curves for the simultaneous showup and standard showup conditions in the mugging

and graffiti videos. Note that, with the data separated by video, there are half of the number of data points that we had planned to collect in each of the simultaneous and standard showup conditions. As such, we use this analysis to check whether the trend was the same for both videos. First, considering the mugging video, it is clear from Figure 1C that performance was close to chance. There was also no evidence of a simultaneous showup advantage because the results trended in the opposite direction, an outcome not predicted by either account. Despite this unexpected trend, the *p*AUC (specificity = .55) for the simultaneous showup (*p*AUC = .14, 95% CI [.11, .16]) did not differ significantly from the *p*AUC for the standard showup, *p*AUC = .17, 95% CI [.14, .20], $D = 1.65$, $p = .098$. Figure 1D shows that the same pattern is observed when the full ROCs are plotted. Again, the trend for an unexpected standard showup advantage is evident in the full ROC curves (Figure 1D), but the AUC for the simultaneous showup (AUC = 0.57, 95% CI [0.52, 0.62] did not differ significantly from the AUC for the standard showup, AUC = 0.64, 95% CI [0.59, 0.69], $D = 1.96$, $p = .05$.

Yet, a very different story emerged when we considered performance in the graffiti video. It is clear from the *p*ROC curves in Figure 1E that subjects' discriminability was better in the simultaneous showup than the standard showup. The *p*AUC (specificity = .82) for the simultaneous showup (*p*AUC = .08, 95% CI [.07, .10]) was greater than the *p*AUC for the standard showup, *p*AUC = .06, 95% CI [0.04, 0.08], $D = 2.01$, $p = .045$. As predicted by the diagnostic-feature-detection theory, presenting similar-looking faces around the graffiti suspect enhanced subjects' collective ability to tell the difference between the real perpetrator and the innocent suspect. Interestingly, Figure 1F shows that the simultaneous showup full ROC is higher than the standard showup ROC on the left side of the graph (i.e., the Yes responses), but the curves come together and then overlap on the right side of the graph where the No responses are included. The AUC for the simultaneous showup (AUC = 0.81, 95% CI [0.79, 0.85]) did not differ significantly from the AUC for the standard showup, AUC = 0.79, 95% CI [0.75, 0.82], $D = 1.06$, $p > .250$. These results show that the predicted effect for choosers, which was evident in the *p*AUC analysis of the Yes response data, is reduced when the non-chooser data are included in the full ROC analysis.

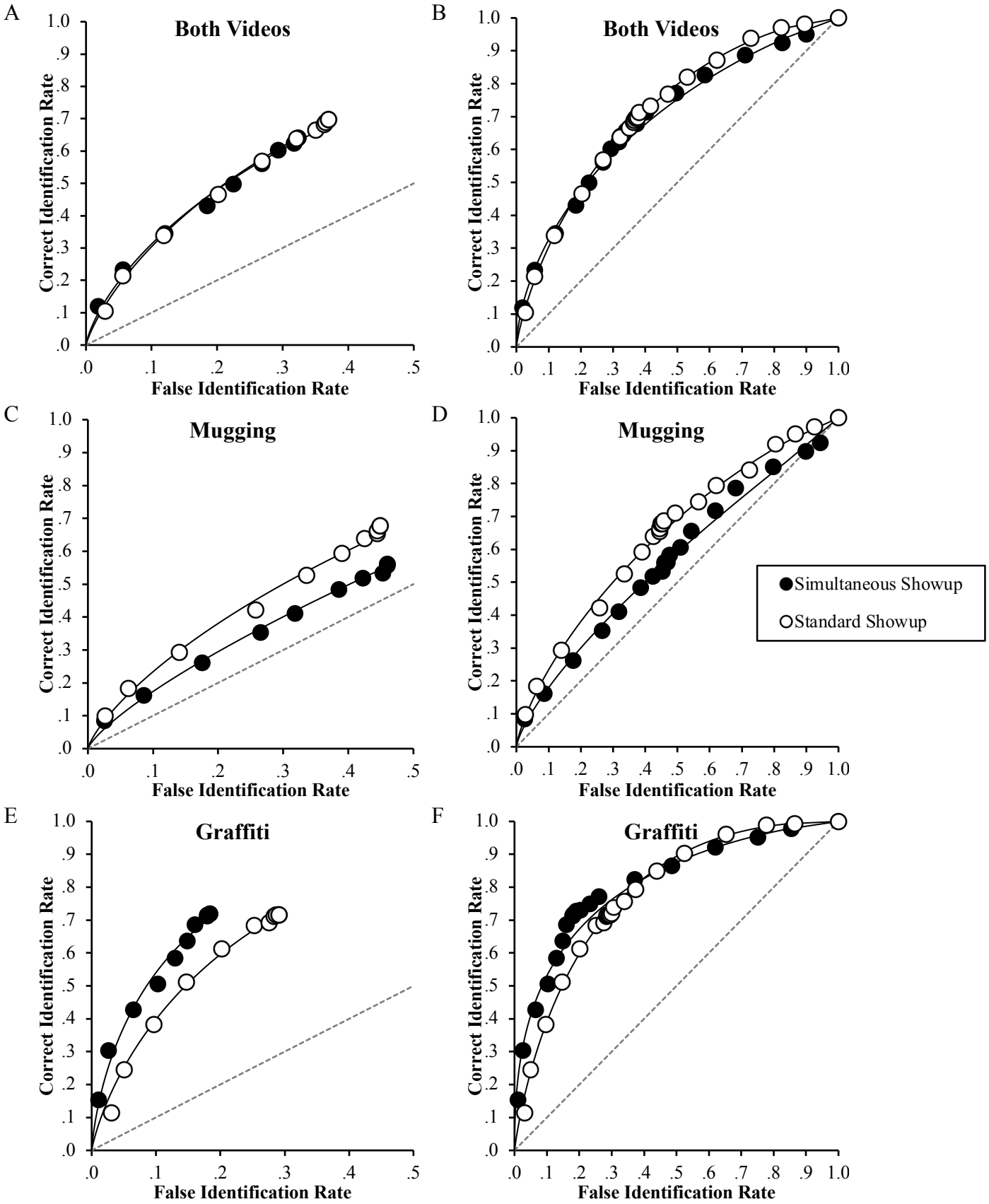

Figure 1. Receiver Operating Characteristic (ROC) curves for the simultaneous showup and the standard showup conditions for (A) choosers and (B) both choosers and non-choosers collapsed over both mock crime videos; for (C) choosers and (D) both choosers and non-choosers in the mugging video; and for (E) choosers and (F) both choosers and non-choosers in the graffiti video in Experiment 1. The lines of best fit were constructed using parameters estimated by the best-fitting unequal-variance signal-detection model. The dashed lines represent chance-level performance.

These results raise two questions. First, why did we find evidence of the predicted simultaneous showup advantage when subjects had watched the graffiti video, but not when they had watched the mugging video? And why were subjects so poor at identifying the real mugging perpetrator? One possibility is that our mugging simultaneous showup was unfair in that our innocent suspect stood out because he was more similar to the participant's *memory* of the perpetrator than the other fillers. As noted earlier, under such conditions, the diagnostic-feature-detection model does not predict a simultaneous showup advantage (e.g., Colloff et al., 2016). Our mock-witness pilot test only serves to illustrate that the innocent suspect was not perceptually distinct from the fillers, based on the description of the perpetrator. To check whether our innocent suspect was more similar to the participant's *memory* of the perpetrator than the other fillers (the key consideration), we examined the identification decisions made to the same faces used as target-absent simultaneous lineups in a different study. After watching the mugging video, 45% of subjects who made an identification selected our innocent suspect from the lineup (far higher than the expected 17% for a fair target-absent lineup)[1]. After watching the graffiti video, 20% of subjects who made an identification selected our innocent suspect from the lineup (much closer to the expected 17%). Therefore, the faces we used for our mugging simultaneous showup did not provide a sound test of our hypothesis. We addressed this issue in Experiment 2.

Second, why might the predicted effect that we observed in the Yes responses (*p*AUC analysis) in the graffiti video be reduced when No responses are included in the full AUC analysis? We did not predict a priori that the findings would differ for choosers (Yes responses) and non-choosers (No responses). To date, ROC research comparing lineups and showups has focused on those subjects who made a positive identification and, as such, the theoretical debate has been concerned with accounting for differences across the two identification procedures in choosers. Nevertheless, it is perhaps unsurprising that the performance of non-choosers differed significantly from choosers, given that choosers and non-choosers have been found to differ in other ways in eyewitness identification tasks (e.g., Brewer & Wells, 2006; Sauer, Brewer, Zweck, & Weber, 2010; Sporer, Penrod, Read, & Cutler, 1995). Although analysis of Yes responses (chooser data) is the key test of the prediction made by the diagnostic-feature-detection theory, it is important to also understand how choosers and non-choosers might differ. As such, in Experiment 2, we also wanted to examine whether the difference between choosers and non-choosers replicated.

[1] Given that no lineup is *perfectly* fair, the expectation is that a randomly selected designated innocent suspect would, if anything, be chosen less than 17% of the time because the odds are only 1 in 6 that the designated innocent suspect will be the most familiar person in the lineup (Palmer, Brewer, Weber, & Nagesh, 2013).

Experiment 2

The diagnostic-feature-detection model predicts that presenting similar-looking faces alongside the suspect (a simultaneous showup) enhances witnesses' ability to discriminate between innocent and guilty suspects compared to presenting the suspect alone (a standard showup). In Experiment 1, we found evidence of the predicted simultaneous showup advantage in choosers using the fair graffiti fillers, but not using the unfair mugging fillers. In Experiment 2, we conducted a fixed-*N* replication ($N = 1{,}000$) using fair fillers for both videos.

Method

Design

We used the same design as Experiment 1.

Subjects

The subjects were 1,076 undergraduates from UCSD who received course credit for participating in the experiment. None of the subjects who had participated in Experiment 1 participated in Experiment 2. We excluded 73 people (6.78% in total) who had completed the experiment more than once ($n = 42$), experienced technical difficulties while watching the video ($n = 5$), or incorrectly answered an attention check question on the content of the video ($n = 26$). This resulted in a final sample of 1,003. Table 1 shows a demographic breakdown of the sample.

Materials

We used the same videos as in Experiment 1.

Showups

For the graffiti perpetrator, we used the same innocent suspect and fillers as in Experiment 1. As noted earlier, after watching the graffiti video in another study, 20% of subjects who made an identification selected our innocent suspect from a simultaneous lineup (close to the 17% expected by chance). However, we adjusted the stimuli for the mugging perpetrator. We removed the innocent suspect and another filler who we judged to be very similar-looking to the perpetrator. To replace these, we randomly selected 2 new fillers from Colloff et al.'s (2016) filler pool. We then randomly selected one of these fillers to serve as the innocent suspect. The diagnostic-feature-detection theory does not predict a simultaneous showup advantage when the innocent suspect is more similar the witness's memory of the perpetrator than the other faces presented (i.e., when the simultaneous showup is unfair). To confirm that our new target-absent simultaneous showup provided a good test of the diagnostic-feature-detection theory, we examined the identification decisions made to the same faces used as target-absent simultaneous lineups in a different study. After watching the mugging video, 16% of subjects who made an identification selected our innocent suspect from the lineup (close to the 17% expected by chance). This illustrates that, our new mugging innocent

suspect was not more similar to the witness's memory of the perpetrator than the other faces presented and, as such, our new mugging stimuli should provide a good test of our hypothesis.

Following convention in the eyewitness literature, we also conducted a standard mock-witness test to determine whether the fillers and innocent suspect in our new mugging simultaneous showup were plausible alternatives to the perpetrator. We provided a new group of mock-witnesses with a description of the perpetrator, and either a target-present or target-absent simultaneous lineup. Again, we refer to these as lineups, because the mock-witnesses were not aware of whom the suspect was and were allowed to pick any face. Forty different mock-witnesses viewed each lineup (total $N = 80$). Tredoux's E' was 3.57 (95% CI [2.85, 4.77]) for the target-present lineup, and was 3.52 (95% CI [2.94, 4.40]) for the target-absent lineup. Again, this indicates that in each lineup there were approximately 4 members who were viable alternatives from which the witness might choose. The perpetrator and innocent suspect were both chosen by 32.5% of the mock-witnesses. Together, these results are similar to previous field (Valentine & Heaton ,1999) and laboratory work (Horry, Palmer, & Brewer, 2012) and suggests that our new mugging simultaneous showup members fit the description of the perpetrator.

Procedure

We used the same procedure as Experiment 1.

Results & Discussion

Recall that our aim to was to determine whether presenting similar-looking faces alongside the suspect (a simultaneous showup) enhances witnesses' ability to discriminate between innocent and guilty suspects compared to presenting the suspect alone (a standard showup). Again, we addressed this question by constructing ROC curves and computing *p*AUC and AUC to measure empirical discriminability. Now that both the mugging and graffiti stimuli were fair and fit-for-purpose, we also fit a signal-detection process model to our data to compare underlying theoretical discriminability (d' or d_a) in the simultaneous showup and standard showup conditions. Finally, we constructed confidence accuracy characteristic curves, because little research has considered the relationship between confidence and accuracy in showups. Our data are freely available (https://osf.io/kyvqa/).

ROC Analysis

Collapsed over both videos. Figure 2A shows the *p*ROC curves for the simultaneous showup and standard showup, collapsed over the two mock-crime videos. As predicted by the diagnostic-feature-detection theory, the simultaneous showup enhanced discriminability compared to the standard showup. Indeed, the *p*AUC (specificity = .87) for the simultaneous showup (*p*AUC = .06, 95% CI [.05, .07]) was significantly greater than the *p*AUC for the standard showup, *p*AUC = .04, 95% CI [.03, .05], $D = 2.61$, $p = .009$. Interestingly, Figure 2B shows that the simultaneous showup

full ROC is higher than the standard showup ROC on the left side of the graph (i.e., the Yes responses), but the curves come together and overlap on the right side of the graph where the No responses are included. The AUC for the simultaneous showup (AUC = 0.83, 95% CI [0.80, 0.85]) did not differ significantly from the AUC for the standard showup, AUC = 0.80, 95% CI [0.77, 0.83], $D = 1.33$, $p = .183$. These results show that the predicted effect for choosers, which was evident in the chooser data, is reduced when the non-chooser data are included in the full ROC analysis.

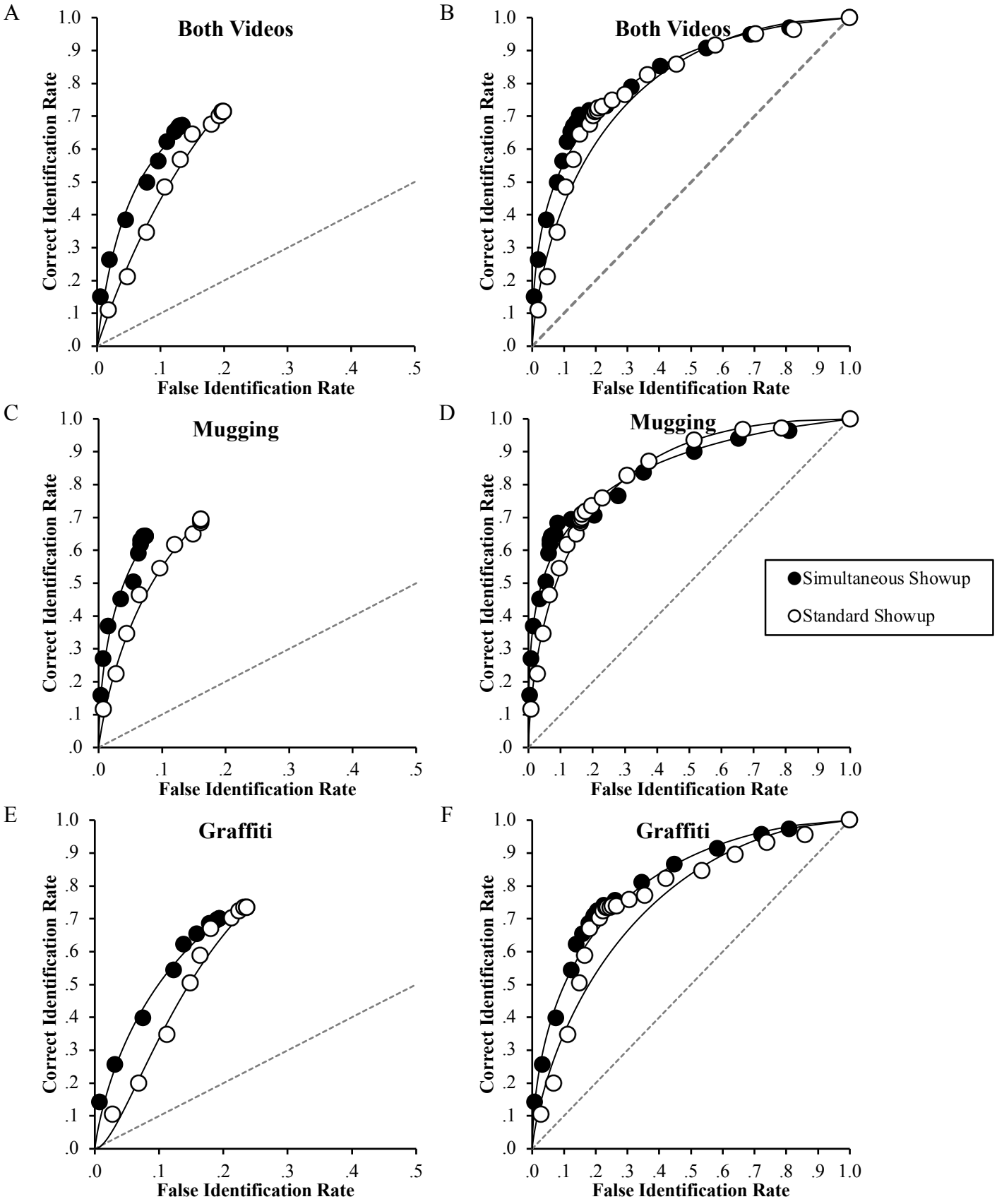

Figure 2. Receiver Operating Characteristic (ROC) curves for the simultaneous showup and the standard showup conditions for (A) choosers and (B) both choosers and non-choosers collapsed over both mock crime videos; for (C) choosers and (D) both choosers and non-choosers in the mugging video; and for (E) choosers and (F) both choosers and non-choosers in the graffiti video in Experiment 2. The lines of best fit were constructed using the parameters estimated from (A) the *best fitting reduced model,* (B) the *constrained choosers and non-choosers model* reported in Appendix A, or (C-F) the parameters estimated by the best-fitting unequal-variance signal-detection model. The dashed lines represent chance-level performance.

Separated by video. Next, we considered performance in both videos separately. It was immediately apparent that, unlike in Experiment 1, identification accuracy for both the mugging video (simultaneous showup $d' = 1.81$, standard showup $d' = 1.50$) and the graffiti video (simultaneous showup $d' = 1.39$, standard showup $d' = 1.34$) was well above chance. This suggests that our innocent suspects were not more similar to the perpetrator than the other filler faces and indicates that both videos now provided a sound test of our hypothesis.

Figure 2C-F show the partial and full ROC curves for the simultaneous showup and standard showup conditions in the mugging and graffiti videos. Again, note that we use this analysis to check whether the trend was the same for both videos, because with the data separated by video, there are half of the number of data points that we had planned to collect in each of the simultaneous and standard showup conditions. The trend was the same in both videos. In the mugging video, the *p*AUC (specificity = .92) for the simultaneous showup (*p*AUC = .04, 95% CI [.03, .05]) was greater than the *p*AUC for the standard showup, *p*AUC = .02, 95% CI [.02, .03], but this was not statistically significant, $D = 1.89$, $p = .059$. In the graffiti video, the *p*AUC (specificity = .81) for the simultaneous showup (*p*AUC = .09, 95% CI [.07, .10] was greater than the *p*AUC for the standard showup, *p*AUC = .06, 95% CI [.04, .08], but again this was not statistically significant, $D = 1.78$, $p = .074$. Taken together, these results fit with the prediction of the diagnostic-feature-detection theory. Presenting similar-looking faces alongside the suspect was enough to improve subjects' ability to discriminate between innocent and guilty suspects, and the predicted trend was apparent in both mock-crime videos.

Additionally, in both the mugging and graffiti videos the simultaneous showup and standard showup curves begin to overlap when No responses are included in the full ROCs (Figure 2D,E). In the mugging video, the AUC for the simultaneous showup (AUC = 0.84, 95% CI [0.81, 0.88]) did not differ significantly from the AUC for the standard showup, AUC = 0.84, 95% CI [0.81, 0.87], $D = 0.04$, $p > .250$. In the graffiti video, the AUC for the simultaneous showup (AUC = 0.81, 95% CI [0.77, 0.85]) did not differ significantly from the AUC for the standard showup, AUC = 0.76, 95% CI [0.77, 0.81], $D = 1.66$, $p = .10$. Again, these results show that the predicted effect for choosers, which was evident in the chooser data, is reduced when the non-chooser data are included in the full ROC analysis.

The results of an ROC analysis based on an atheoretical measure like *p*AUC need not agree with results based on a theoretical measure like d' (or d_a) obtained by fitting a theoretical model to the same data (Lampinen, 2016; Rotello & Chen, 2016). In fact, the two measures can go in opposite directions (see Wilson et al., in press; Wixted & Mickes, 2018) even though they usually agree. To further confirm our findings, we fit a signal-detection process model to our data (Wixted & Mickes,

2014). Recall that only the diagnostic-feature-detection theory—and not the filler siphoning theory—predicts that presenting similar-looking faces in a simultaneous showup will improve subjects' underlying theoretical discriminability. Our model fitting analyses are presented in Appendix A and agree with the *p*AUC analyses. In short, for choosers discriminability was significantly better in the simultaneous showup (d_a = 1.54) than in the standard showup (d_a = 1.45). For non-choosers, discriminability was the same across the showup conditions (d' = 1.21).

We did not predict a priori that the performance of non-choosers would differ significantly from choosers (though for a similar pattern, see Colloff, Wade, Strange, & Wixted, 2018). We can think of two possible post hoc interpretations. One possibility is that subjects in the non-chooser group did not encode the perpetrator very well. Both the chooser and non-chooser groups are heterogeneous groups of people, because, in each group, there are some people who encoded the face in the mock-crime video well, and other people who did not encode the face well. The heterogeneous group that had their decision criteria set in such a way that they made a positive identification (i.e., the chooser group), have an average d_a that is fairly high. The heterogeneous group that had their decision criteria set in such a way that they did not make a positive identification (i.e., the non-chooser group), have an average d' that is fairly low. Although both the chooser and non-chooser groups are a mixture of individuals, on average, non-choosers are less able to discriminate between innocent and guilty suspects, possibly because they encoded the face in the mock-crime video less strongly than choosers. As such, the predicted effect—that presenting similar-looking faces around the suspect will increase ability to discriminate between innocent and guilty suspects—may not occur in non-choosers, because, generally speaking, these are the people who did not encode the perpetrator's face very well to begin with.

Another possibility is that the predicted effect—that presenting similar-looking faces around the suspect will increase ability to discriminate between innocent and guilty suspects—*does* occur in non-choosers, but the expected difference is obscured by decision noise. That is, perhaps criterion variance increases as you move from conservative decisions (i.e., choosers who identified the suspect and rated their confidence as high) to liberal decisions (i.e., non-choosers who did not identify the suspect and rated their confidence as high). This might occur, for example, if some participants misunderstand the confidence scale, using 0 to express a high-confidence "no" decision (i.e., to mean "zero-percent chance this is the perpetrator"), whereas other participants correctly use 100 to express a high-confidence "no" decision. This account would mean that that, on average, non-choosers have greater criterion variance across witnesses than choosers and could therefore explain why an EV model (i.e., $\sigma_{guilty} = \sigma_{innocent} = 1$) applies to the non-choosers, but an UV model in which σ_{guilty} is less than 1 applies to the choosers (see Appendix A).

Future research should investigate which of these two possible explanations—weaker memory or greater criterion variance in non-choosers—best accounts for the difference between choosers and non-choosers. Moreover, research should examine whether choosers and non-choosers also differ significantly on lineup tasks. Although much ROC research has found that simultaneous lineups result in more accurate eyewitness identifications than showups in choosers, it is not yet known whether this is also the case in non-choosers. Therefore, conducting ROC lineup studies in such a way to collect confidence judgements for non-choosers is an important topic of future research. The aim of the current study, however, was to test two theories that have been offered to explain the simultaneous lineup advantage over showups—a discussion which has exclusively focused on choosers. Taken together, the results of the model fitting exercise are concordant with the results of our analyses based on the atheoretical *p*AUC measure. Both suggest that, for choosers, presenting similar-looking faces alongside the suspect enhances witnesses' ability to discriminate between innocent and guilty suspects, despite there being no opportunity for erroneous identifications to be spread across the fillers. This pattern of results is predicted by the diagnostic-feature-detection theory, not the filler siphoning account.

Confidence Accuracy Characteristic Analysis

For each showup condition, we formed a single 21-point confidence scale, ranging from No 100 to No 0 then Yes 0 to Yes 100, and calculated proportion correct separately for each level of confidence (No 100, No 90, No 80, and so forth, following Mickes, 2015). For Yes responses, we calculated proportion correct using the formula: $Y_g / (Y_g + Y_i)$, where Y_g and Y_i are the number of Yes responses to the guilty and innocent suspects, respectively. For No responses, we used the formula: $N_i / (N_g + N_i)$, where N_g and N_i are the number of No responses to the guilty and innocent suspects, respectively. To provide more stable estimates, we binned confidence level into six categories (No 100–90, No 80–70, No 60–0, Yes 0–60, Yes 70–80, Yes 90–100). For each confidence bin, we estimated the standard error using the formula, SE = √[proportion correct * (1 - proportion correct) / *n*], where *n* is the number of observations included in the proportion correct calculation. Figure 3 shows the confidence accuracy characteristic curves for the simultaneous showup and standard showup, collapsed over the two mock-crime videos. Nonoverlapping standard error bars denote reliable differences between the showup conditions (e.g., Sauer, Brewer, Zweck, & Weber, 2010). The standard showup yielded more accurate decisions than the simultaneous showup for low confidence No decisions (No 60-0), but the simultaneous showup yielded more accurate decisions that the standard showup for high confidence Yes decisions (Yes 90-100). Although

confidence better tracked accuracy in the simultaneous showup, the same pattern was apparent for both showup conditions: When confidence was high (No 100-90 and Yes 90-100), subjects were very accurate and when confidence decreased towards uncertainty (No 60-0, Yes 0-60), accuracy decreased to around chance performance. This is interesting because much lineup literature indicates that confidence and accuracy are not related for No decisions (non-choosers; e.g., Sporer et al., 1995). Here, however, in both showup conditions, a high-confidence No decision provides considerable evidence of innocence. This result suggests that highlighting the suspect could offer a way for eyewitness confidence to provide evidence that the police suspect is not the real culprit. We examined this further in Experiment 3, in which we also included a standard lineup condition.

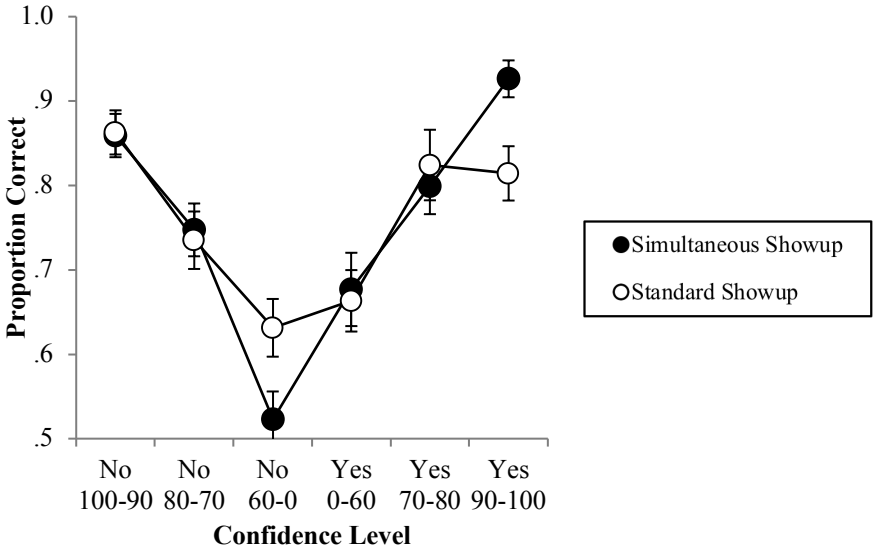

Figure 3. Confidence Accuracy Characteristic plot for the simultaneous showup and standard showup conditions in Experiment 2. Error bars indicate ±1 *SE*.

Experiment 3

In Experiment 1 and 2 we found that, in choosers, presenting similar-looking faces alongside the suspect (a simultaneous showup) enhances witnesses' ability to discriminate between innocent and guilty suspects compared to presenting the suspect alone (a standard showup) when fair fillers are used. This provides evidence for a diagnostic-feature-detection mechanism. What is not yet clear, however, is how people perform on simultaneous showups and standard showups, relative to standard simultaneous lineups.

The diagnostic-feature-detection theory predicts that the ability to discriminate between innocent and guilty suspects (as measured by *p*AUC and fitting a theoretical model) will be better in

both the simultaneous showup and the standard simultaneous lineup compared to the standard showup. This is because both the simultaneous showup and the standard simultaneous lineup allow for comparison of features across multiple faces which enhances witnesses' underlying theoretical ability to discriminate between innocent and guilty suspects, whereas the standard showup does not. The filler siphoning hypothesis, however, does not specifically predict that ability to discriminate between innocent and guilty suspects as measured by fitting a theoretical model—that is, theoretical discriminability—will differ across the three procedures. Yet, filler siphoning theory could be used to argue that the *p*ROC curve for the standard simultaneous lineup will be higher than the *p*ROC curves for both the simultaneous showup and the standard showup (which will lie on top of each other), because the presence of possible alternatives in the lineup siphons some of the incorrect identifications that would have otherwise have landed on the innocent suspect. According to filler siphoning theory, this process could have the net effect of raising the *p*ROC curve for lineups, even though theoretical discriminability is equivalent across all three procedures. To test these hypotheses, we added a standard simultaneous lineup condition in Experiment 3.

In Experiment 3, we also adjusted how the identification task was presented to subjects to increase the ecological validity and generalizability of our results. First, in Experiments 1 and 2 we asked subjects to rate their confidence that they would be able to recognize the perpetrator; In Experiment 3, we omitted this question because (a) the resultant data do not test the two theoretical accounts of the lineup advantage and (b) recent research suggests that pre-identification confidence judgments may influence post-identification confidence judgments (Bednarz, Carlson, Carlson, Wooten, & Young, 2016). Second, in Experiments 1 and 2, the photo(s) were displayed for 10 s before a red border appeared around the suspect. It was only after 10 s that subjects were told about their task and were able to make a yes/no identification decision. We used this delayed procedure in an effort to ensure that subjects attended to the additional faces in the simultaneous showup condition. The 10 s delay, however, differs from how standard showups and lineups are often conducted in laboratory studies and in the real world, because subjects and witnesses are usually instructed on their task before they are presented with the identification procedure, and are able to make a decision in their own time. Therefore, in Experiment 3, we told subjects about their task before the images were presented and we removed the 10 s delay. Third, in Experiments 1 and 2, subjects were instructed that the role of the additional faces in the simultaneous showup was to help them decide whether the suspect was the person that committed the crime. These instructions are vague (how, exactly, will those faces help in the decision?), so in Experiment 3 we used instructions that specified the way in which the additional faces in the simultaneous showup might help. We instructed subjects that the role of the faces was to show what an innocent suspect might look like.

Finally, in Experiments 1 and 2, we used pre-designated faces to serve as the innocent suspects and fillers. While this permitted the greatest amount of experimental control, it meant that our results were limited to just two guilty-innocent suspect pairs and a small subset of filler faces. In Experiment 3, we randomly generated the innocent suspect and filler faces for each subject from pools of faces to ameliorate the problems associated with using pre-designated innocent suspect and filler faces and assessing lineup fairness.

Method

Design

We used a 3 (presentation: simultaneous showup, standard showup, standard simultaneous lineup) × 2 (video: mugging, graffiti) × 2 (target: present, absent) mixed design, with video and target manipulated within subjects.

Subjects

The subjects were 1,792 undergraduates from UCSD who received course credit for participating in the experiment. None of the subjects who had participated in Experiments 1 and 2 participated in Experiment 3. We excluded 150 people (8% in total) who had completed the experiment more than once (n = 106), experienced technical difficulties while watching the video (n = 7), or incorrectly answered an attention check question on the content of the video (n = 37). This resulted in a final sample of 1,642. Table 1 shows a demographic breakdown of the sample.

Materials

We used the same videos as in Experiments 1 and 2.

Lineups & Showups

We used the pools of fillers compiled by Colloff et al. (2016). Simultaneous showups and standard simultaneous lineups consisted of either the perpetrator and five randomly selected fillers (target-present) or one randomly selected innocent suspect and five randomly selected fillers (target-absent). Standard showups were either a single photo of the perpetrator (target-present), or a randomly selected innocent suspect (target-absent).

Procedure

The procedure was similar to the procedure that we used in Experiment 1 and 2, but we made three changes to how the identification task was presented. First, we omitted the question that asked subjects to rate their confidence that they would be able to recognize the perpetrator. Second, we told subjects what images they would see, before they viewed the images. All subjects were told that they would be asked to decide whether the police suspect was the perpetrator from the video. Those in the simultaneous showup condition were told that a photo of the suspect would be displayed in a red box along with the photos of five other men. They were told: *"The other five men are not suspects; their*

role is to show you what an innocent suspect might look like in a case like this." Those in the standard showup condition were told that they would be presented with a photo of the suspect**.** Those in the lineup condition were told that they would be presented with a lineup of six photos. Third, when the identification task was displayed, we presented subjects with the identification instructions immediately (i.e., we removed the 10 s delay which we used in Experiments 1 and 2) and the images were randomly generated for each subject from pools of images. In the simultaneous showup condition, six faces were displayed simultaneously in two rows of three photos. A thick red border was displayed around the suspect. Subjects were told again: *"The police suspect is highlighted in red. The other five men are not suspects*; *their role is to show you what an innocent suspect might look like in a case like this. The police suspect may or may not be the actual perpetrator."* In the standard showup condition, one photo was displayed. A thick red border was displayed around the image. Subjects were told: "*This is the police suspect. The police suspect may or may not be the actual perpetrator*." Subjects in both showup conditions were asked the same question: *"Is the suspect (highlighted in red) the person who committed the crime?"* and responded by clicking on "Yes" or "No". In the standard lineup condition, six faces were displayed simultaneously in two rows of three photos. Below the faces was an option labeled "Not Present." Subjects were told: *"The lineup above may or may not contain the perpetrator who committed the crime. Please click on the person that you believe is the perpetrator, or choose "Not Present" if you think the perpetrator is not in the lineup."*

After subjects had made an identification decision, the experimental procedure was identical to Experiments 1 and 2. That is, subjects used an 11-point Likert-type scale (0=*completely uncertain* to 100=*completely certain*) to rate their confidence in their decision and answered a question that enabled us to check that they were paying attention ("How many people were in the video?"). The procedure then began again, but, this time, subjects were allocated into the alternate video (mugging or graffiti) and target (present or absent) condition. The order of the video and target conditions was counterbalanced. Finally, at the end of the study, subjects answered a number of demographic questions.

Results & Discussion

Recall that our aim to was to determine whether presenting similar-looking faces alongside the suspect (a simultaneous showup) enhances witnesses' ability to discriminate between innocent and guilty suspects compared to presenting the suspect alone (a standard showup), and if simultaneous showups enhance witnesses' ability to discriminate between innocent and guilty suspects to a similar extent as standard lineups. Again, we addressed this question by calculating *p*AUC and fitting a signal-detection model to our data. Note that, unlike Experiment 1 and 2, Experiment 3 only allows

for an ROC analysis of choosers, because a standard lineup condition only requires that choosers—and not non-choosers—rate their confidence that an individual is the culprit. Non-choosers in the simultaneous lineup condition did not make a confidence judgement to a specific face. Finally, we constructed confidence accuracy characteristic curves to examine the relationship between confidence and accuracy. Our data are freely available (https://osf.io/kyvqa/).

ROC Analysis

Collapsed over both videos. Figure 4A shows the *p*ROC curves for the simultaneous showup, the standard showup and the standard simultaneous lineup conditions, collapsed over the two mock crime videos. Both the simultaneous showup and the standard lineup enhanced subjects' ability to discriminate between innocent and guilty suspects compared to the standard showup. The *p*AUCs (specificity = .91) for the simultaneous showup (*p*AUC = .03, 95% CI [.02, .03], $D = 2.68$, $p = .007$) and the standard lineup (*p*AUC = .03, 95% CI [.03, .04], $D = 4.62$, $p < .001$) were significantly greater than the *p*AUC for the standard showup, *p*AUC = .01, 95% CI [.01, .02]. The simultaneous showup and the standard lineup led to equivalent discriminability; the *p*AUCs did not differ significantly, $D = 1.44$, $p = .150$. This is the pattern of results predicted by the diagnostic-feature-detection theory.

An interesting additional point to note about these data is that, although the curves for the simultaneous showup and the standard lineup fall on top of each other, the curve for the simultaneous showup extends further—reflecting both a higher correct identification rate and false identification rate in this condition. This illustrates that when witnesses know who the suspect is (in the simultaneous showup), they are more likely to choose the suspect compared to when they do not know who the suspect is (in the standard lineup), yet discriminability is unaffected. This is an example of a "suggestive" procedure in action; A suggestive procedure affects response bias, not discriminability.

A possible applied implication of our findings is that the simultaneous showup could be used to extend the simultaneous lineup ROC to the right to yield higher false ID rates, if it is determined that higher false ID rates are desirable, as Smith et al. (in press) suggest will sometimes be the case. Smith et al., recently argued that a utility analysis using a new measure—deviation from perfect performance (DPP)—should be used instead of *p*AUC, because DPP could favour a procedure that falls on a lower ROC if that procedure also yields more liberal responding than the alternative procedure. As an example, they showed that the use of DPP could favour an unfair lineup over a fair lineup even though the unfair lineup yields a higher false identification rate and lower ROC. The problem with DPP, however, is that it is never possible to determine that one procedure is

diagnostically superior to the other, because it depends on subjective values that are unknown, such as the costs and benefits identification decisions, and subjective assumptions about the prior probability of the real culprit being present in the lineup. If the goal is to determine which procedure is diagnostically superior across high false identification rates because high false identification rates are preferred as a matter of policy (an unlikely scenario), instead of using a memory-harming unfair lineup, a better approach would be to extend the simultaneous lineup curve to the right and compare the procedures in the usual way. This can be easily accomplished by making the procedure suggestive by highlighting the suspect (eliciting more liberal responding), and then measuring performance using *p*AUC, which is objective and independent of costs, benefits and prior probabilities.

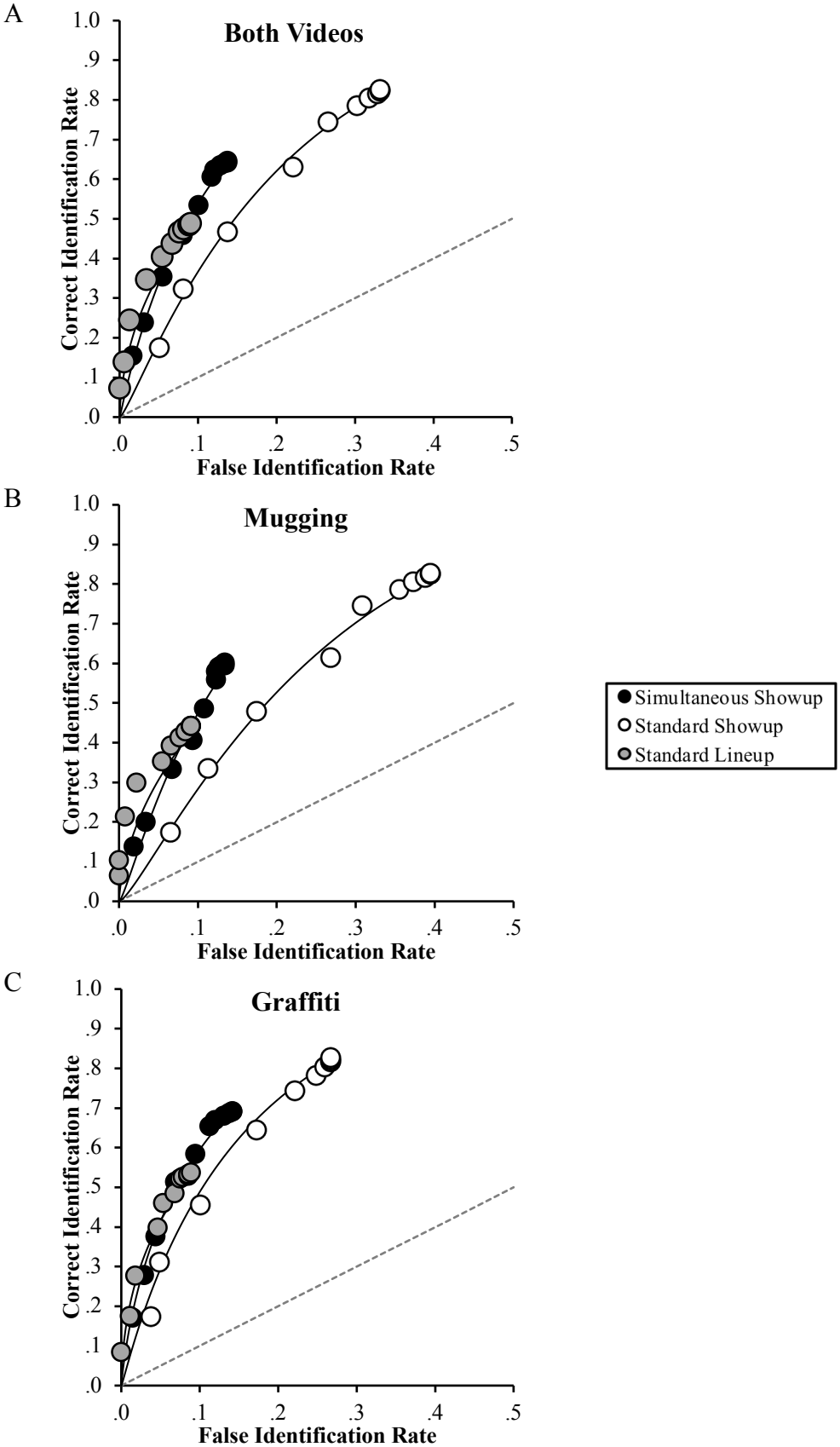

Figure 4. Partial Receiver Operating Characteristic (*p*ROC) curves for the simultaneous showup, standard showup and standard simultaneous lineup conditions (A) collapsed over both mock crime videos, (B) in the mugging video, and (C) in the graffiti video in Experiment 3. The lines of best fit were constructed using the parameters estimated from (A) the *full model* reported in the text, or (B,C) the parameters estimated by the best fitting unequal-variance signal-detection model. The dashed line represents chance-level performance.

Separated by video. Figure 4 shows the *p*ROC curves for the simultaneous showup, the standard showup, and the standard simultaneous lineup conditions in the (B) mugging and (C) graffiti videos. The trend was the same in both videos. In the mugging video, the *p*AUCs (specificity = .91) for the simultaneous showup (*p*AUC = .02, 95% CI [.01, .03], $D = 2.00$, $p = .046$) and the standard lineup (*p*AUC = .03, 95% CI [.02, .04], $D = 4.40$, $p < .001$) were greater than the *p*AUC for the standard showup, *p*AUC = .01, 95% CI [.01, .02]. The *p*AUCs for the simultaneous showup and standard lineup did not differ significantly, $D = 1.63$, $p = .103$. In the graffiti video, the *p*AUCs (specificity = .91) for the simultaneous showup (*p*AUC = .03, 95% CI [.02, .04], $D = 1.61$, $p = .108$) and standard lineup (*p*AUC = .04, 95% CI [.03, .04], $D = 2.39$, $p = .017$) were greater than the *p*AUC for the standard showup (*p*AUC = .02, 95% CI [.01, .03]), though only the difference between the standard lineup and the standard showup was statistically significant. The *p*AUCs for the simultaneous showup and standard lineup did not differ significantly, $D = 0.48$, $p = .634$.

Again, we fit a signal-detection model to further confirm our findings (see Appendix B). Our model-fitting analyses largely agreed with the *p*AUC analyses. Namely, underlying theoretical discriminability was better in the simultaneous showup ($d_a = 1.56$) and the standard simultaneous lineup ($d_a = 1.25$) than in the standard showup ($d_a = 1.23$). Taken together, these results fit with the prediction of the diagnostic-feature-detection model. Presenting similar-looking faces alongside the suspect in a simultaneous showup enhanced subjects' ability to discriminate between innocent and guilty suspects compared to presenting the suspect alone. Indeed, presenting similar-looking faces alongside the suspect in a simultaneous showup enhanced ability to discriminate between innocent and guilty suspects to the at least the same degree as a standard simultaneous lineup.

Confidence Accuracy Characteristic Analysis

Figure 5 shows the confidence accuracy characteristic curves for the simultaneous showup, standard showup, and standard simultaneous lineup, collapsed over the two mock-crime videos. The standard showup yielded more accurate decisions than the simultaneous showup for No decisions (No 100-90, 80-70), but the simultaneous showup yielded more accurate decisions that the standard showup for Yes decisions (Yes 70-80, Yes 90-100). Moreover, the standard simultaneous lineup yielded the most accurate high confidence Yes decisions (Yes 90-100), but the least accurate high confidence No decisions (No100-90). This fits with previous lineup literature, which consistently finds that confidence is related to accuracy in choosers (Yes responses), but not in non-choosers (No responses; Sporer et al., 1995). Interestingly, Figure 5 shows that in showups, confidence tracks accuracy in non-choosers; A high-confidence No decision from the simultaneous and standard showups, unlike the standard lineup, provides considerable evidence of innocence. It seems sensible

to further explore this phenomenon using a procedure that could be more easily employed in practice. For example, a witness could first view a standard simultaneous lineup. If the witness says "No, the real culprit is not present", then the police suspect could be highlighted and the witness could be asked to rate their confidence that the police suspect is not the culprit. Such a procedure could dramatically increase the quantity of information that is collected during an identification task by eliciting reliable evidence of innocence for non-chooser lineup identification decisions.

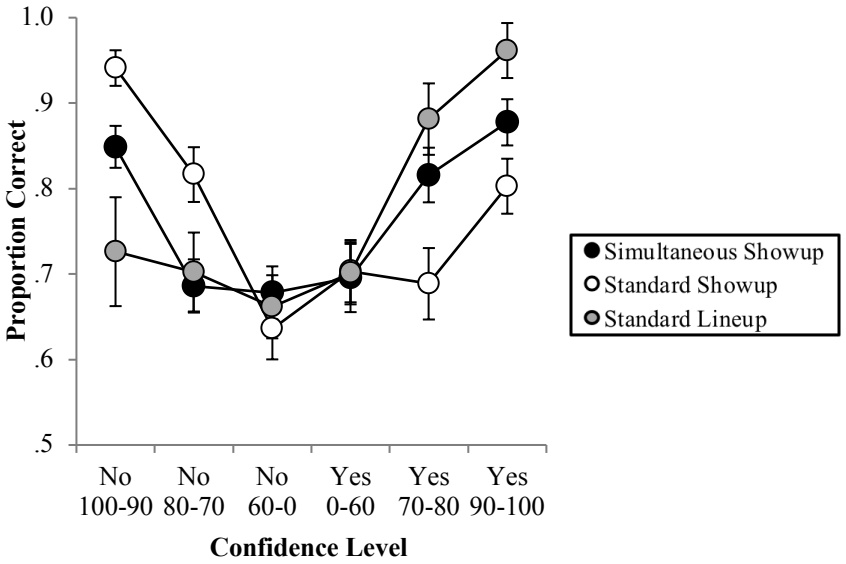

Figure 5. Confidence Accuracy Characteristic plot for the simultaneous showup, standard showup and standard simultaneous lineup conditions in Experiment 3. Error bars indicate ±1 *SE*.

General Discussion

We investigated why simultaneous lineups aid identification performance more than showups. When fair identification procedures were used, we found that simply presenting similar-looking faces alongside the suspect (like a simultaneous lineup) enhanced subjects' ability to discriminate between innocent and guilty suspects compared to presenting the suspect alone (a showup). Indeed, presenting similar-looking faces alongside the suspect enhanced subjects' ability to discriminate between innocent and guilty suspects to the same extent as a standard simultaneous lineup.

Many previous studies have shown that simultaneous lineups result in more accurate identifications than showups (Clark & Godfrey, 2009; Gronlund et al., 2012; Key et al., 2015; Lindsay et al., 1997; Steblay et al., 2003; Wetmore et al., 2015). But we found that this lineup advantage exists when the suspect is known and there is no opportunity for erroneous identifications to be spread across the other lineup members. This pattern of results is theoretically interesting,

because it is only predicted by one account of eyewitness identification performance—the diagnostic-feature-detection theory (Wixted & Mickes, 2014). The theory suggests that simultaneous lineups enhance discriminability more than showups because lineups afford witnesses the opportunity to discern which facial features are shared by all lineup members and then discount these shared features from the identification decision. Because showups only contain the suspect, there is no opportunity for witnesses to learn which facial features may and may not be useful for the identification decision. According to the diagnostic-feature-detection theory, presenting similar-looking faces alongside the suspect is enough to enhance witnesses' ability to tell the difference between real perpetrators and innocent suspects. Indeed, this is what we found.

Critically, filler siphoning theory does not predict and therefore cannot account for our findings (Wells, 2001). Filler siphoning theory holds that simultaneous lineups result in more accurate identifications than showups because some of the incorrect identifications in lineups land on the fillers instead of the innocent suspect. Because showups do not contain fillers, all of the incorrect identifications land on the innocent suspect (Wells, Smalarz, & Smith, 2015; Wells, Smith, & Smalarz, 2015). According to filler siphoning theory, the opportunity to spread erroneous identifications over multiple faces is the key to the lineup ROC advantage. We removed this opportunity in our study and still found that similar-looking faces enhanced subjects' ability to sort guilty and innocent suspects into their appropriate categories.

Conceivably, some form of "mental filler siphoning" occurred in our simultaneous showup condition, whereby participants mentally chose another face and did not choose the suspect as a result. However, this seems unlikely for several reasons. First, filler siphoning occurs when a witness believes that a filler is more likely to be the perpetrator than the suspect and chooses the filler for that reason. However, in the simultaneous showup condition, subjects were explicitly told that the other 5 faces were not suspects, and they were asked to make a yes/no decision about only the suspect. Second, mental filler siphoning would have been more likely to occur in Experiments 1 and 2 than in Experiment 3 because in Experiments 1 and 2 there was a 10 s delay in which subjects could have mentally selected an alternative face from the display before the identification instructions appeared. We removed this delay in Experiment 3, yet the magnitude of the simultaneous showup advantage over standard showups as measured by D in our pAUC analysis was similar in all three experiments (Experiment 1 fair graffiti fillers $D = 2.01$; Experiment 2 collapsed over both mock crimes, $D = 2.61$;

Experiment 3 collapsed over both mock crimes, $D = 2.68$; pairwise comparisons using one-tailed z tests, all ps > .250).[2]

Finally, and perhaps most importantly, filler siphoning cannot explain how the simultaneous showup condition enhanced subjects' underlying discriminability when the data are analyzed by fitting a theoretical model—at least not when fitting the signal-detection model that we used to interpret the data (Wixted & Mickes, 2014)—because the filler siphoning account does not predict a change in underlying discriminability. In addition, the signal-detection model naturally *predicts* (and therefore explains) the phenomenon of filler siphoning. In a signal-detection model, the inclusion of fillers in a lineup is theoretically represented by a filler distribution. Although the *mean* of the filler distribution is lower than the mean of the target distribution, some fillers in a target-present lineup will generate a strong memory match signal. Sometimes, the signal generated by the strongest filler will exceed the memory-match signal generated by the target, leading to filler siphoning. A similar phenomenon occurs in target-absent lineups (i.e., the memory-match signal of the strongest filler must sometimes exceed the memory-match signal generated by the innocent suspect). As such, the effect of filler siphoning is akin to a change in response bias—increased filler siphoning results in a decrease in both guilty and innocent suspect identifications, just like increasingly conservative responding results in a decrease in both guilty and innocent suspect identifications (see Colloff, Wade, Strange, & Wixted, 2018). The key point is that even though the signal-detection model can naturally account for filler identifications (i.e., it can account for filler siphoning), that filler-siphoning phenomenon alone is not sufficient to explain the differences we observed in the data. In addition to accounting for filler siphoning, our model fits also required a discriminability advantage in the simultaneous showup condition compared to the standard showup condition to adequately fit the data. Filler siphoning theory does not make a prediction about underlying theoretical discriminability.

Although it cannot explain our simultaneous showup data, filler siphoning *can* predict that the presence of fillers in a simultaneous lineup will raise the simultaneous lineup *p*ROC above the standard showup *p*ROC when instantiated in a quantitative model. Wetmore, McAdoo, Gronlund, and Neuschatz (2017) recently conducted simulations using the WITNESS model and found that,

[2] Similarly, mental filler siphoning would have been more likely to occur on the first identification task than the second identification task in Experiment 2 because, presumably, by the second identification task subjects would be aware that one face would be highlighted after the 10 s delay and they would not be permitted to choose an alternative face. Again, additional ROC analyses revealed that the magnitude of the simultaneous showup advantage over standard showups was similar on the first identification task and the second identification task. In the first ID task $D = 1.78$ (specificity = .87; simultaneous showup: *p*AUC = .05, 95% CI [.04, .06]; standard showup: *p*AUC = .04, 95% CI [.03, .05], and in the second ID task $D = 1.89$ (specificity = .87; simultaneous showup: *p*AUC = .06, 95% CI [.05, .07]; standard showup: *p*AUC = .04, 95% CI [.03, .06]). D was similar across the first and second ID task, $z = 0.11$, $p > .250$ (one tailed).

although filler siphoning can raise the *p*ROC curve for simultaneous lineups compared to showups, the size of filler siphoning effect was not large enough to explain the size of the effects observed in empirical studies. This provides further evidence that filler siphoning alone is not a sufficient account of the empirical data. The simulations conducted by Wetmore et al. (2017) did generate another interesting possibility—that criterial variability could *lower* the *p*ROC for standard showups compared to simultaneous lineups (see also Smith, Wells, Lindsay, & Penrod, 2017). Put another way, fillers in a lineup may mitigate the negative effect of criterial variability. In Experiment 3, we found that the *p*ROCs for the standard lineup and the simultaneous showup overlapped and were both higher than the *p*ROC for the standard showup. The criterial variance account does predict (and therefore could account for) the higher *p*ROC for standard lineups compared to standard showups. However, it does not predict a higher *p*ROC for simultaneous showups compared to standard showups. A post hoc interpretation of our results in terms of the criterial variance hypothesis could be that viewing filler faces (not the opportunity to identify one of the filler faces) was enough to mitigate the negative effect of criterial variability. Although this could be an interesting avenue for future research, it seems fair to say that there is no obvious reason for the criterial variance hypothesis to anticipate this effect. Conversely, the diagnostic feature-detection theory a priori predicts both the standard lineup advantage and the simultaneous showup advantage that we observed. As such, currently, the diagnostic-feature-detection theory is the only theory that comfortably predicts our findings.

Why is this important? It is important to understand the underlying psychological mechanisms because this knowledge can be used to further enhance eyewitness accuracy. Indeed, the diagnostic-feature-detection model predicts that comparison of facial features across lineup members aids identification performance. We used this prediction to improve subjects' discriminability on a suggestive showup procedure so that it was similar to subjects' ability to discriminate between innocent and guilty suspects from a standard simultaneous lineup. When theoretical models are refined, it is easy to see how they could be applied to develop or modify identification techniques (Gronlund, Mickes, Wixted, & Clark, 2015; Wixted & Mickes, 2014). Our results suggest that it might be possible to improve showup accuracy in the field simply by presenting witnesses with a simultaneous set of description-matched faces before asking whether or not the suspect is the perpetrator. Practically, this is important because showups are the most commonly used identification procedure in England and in the US (Davis, Valentine, Memon, & Roberts, 2015; Gonzalez, Ellsworth, & Pembroke, 1993). Nevertheless, in practice, policymakers who are concerned with keeping the false alarm rate of innocent suspects low might favour a standard simultaneous lineup because it is less suggestive and elicits more conservative responding (i.e.,

fewer suspect IDs) than the simultaneous showup. Suspect (chooser) IDs made with high confidence were also more reliable in the standard simultaneous lineup than the simultaneous showup. Interestingly, however, our results suggest that the simultaneous showup procedure could dramatically increase the informativeness of non-chooser lineup identification decisions, because it elicits reliable evidence of innocence. Future research should assess the utility of a procedure in which a witness is first shown a standard simultaneous lineup, but if they answer "not present", the suspect is highlighted (like in a simultaneous showup) and evidence of innocence collected. Finally, our results illustrate how the simultaneous showup procedure could be used to extend the simultaneous lineup ROC to the right to incorporate higher false ID rates, in cases where the goal is to determine which lineup procedure is diagnostically superior across high false identification rates (cf. Smith et al., in press).

We should note, however, that we did not observe a simultaneous showup advantage in one stimulus set in Experiment 1. This is likely due to the innocent suspect looking more like the actual perpetrator than the other fillers, which may explain not only the failure to find a simultaneous showup advantage but also the very poor discriminative performance observed in that condition. Because the diagnostic-feature-detection hypothesis does not predict a simultaneous lineup advantage when the lineup is unfair, the faces we used in our mugging simultaneous showup did not provide a sound test of our hypothesis (Colloff et al., 2016). Nevertheless, the simultaneous showup advantage was evident in both stimulus sets in Experiment 2 and when the innocent suspect and filler faces were randomly generated for each subject from a pool of faces in Experiment 3. This illustrates that our findings are not driven by a particular guilty and innocent suspect pair, nor the specific encoding and test conditions of a particular stimulus event (e.g. Brewer, Keast, & Sauer, 2010; Lindsay, Read, & Sharma, 1998). Moreover, the simultaneous showup advantage was evident, regardless of the change in instructions from Experiment 2 to 3. The enhanced discriminability afforded by presenting similar-looking faces alongside the perpetrator appears to be a general phenomenon. At the very least, we can conclude that the phenomenon appears to hold across a variety of testing conditions when the perpetrator has a distinctive feature and when fair fillers are used. We have no a priori reason to predict that the pattern of results would be different when the culprit does not have a distinctive feature, though this is, of course, an empirical question that requires testing. The modelling exercise further confirms that the results of the theory-free, objective ROC analyses, by-and-large, map onto measures of underlying theoretical discriminability (Rotello & Chen, 2016).

While we can be reasonably confident that our simultaneous showup advantage is a reliable effect in our studies, it is critical to highlight that our findings do not *disprove* the existence of filler

siphoning in lineup tasks. Filler siphoning occurs in real world identification decision; witnesses often choose fillers, and that phenomenon is predicted by any signal-detection model of lineup performance. Although there is no a priori reason to expect that filler siphoning, per se, would have the effect of elevating the ROC, the simulations reported by Wetmore et al. (2017) suggest that a small effect in that direction can happen. Our findings suggest that there is more to the story than that. Specifically, our results show, for the first time, that at least some (if not all) of the discriminability advantage in simultaneous lineups is due to the comparison of multiple faces as predicted by the diagnostic-feature-detection account. Future research should continue to examine the differences in performance between choosers and non-choosers to enhance our theoretical understanding of the decision process.

So, in sum, why are simultaneous lineups better than showups? One reason is that the presence of similar-looking faces alongside the suspect enhance people's collective ability to discriminate between the real culprit and an innocent suspect, as predicted by the diagnostic-feature-detection theory.

References

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

Bednarz, J., Carlson, C., Carlson, M., Wooten, A., & Young, D. (2016, March). *Eyewitness confidence and accuracy: An evaluation of pre- versus post-lineup confidence.* Poster presented at the meeting of American Psychology-Law Society, Atlanta, GA.

Brewer, N., Keast, A., & Sauer, J. D. (2010). Children's eyewitness identification performance: Effects of a Not Sure response option and accuracy motivation. *Legal and Criminological Psychology*, *15*, 261–277. http://dx.doi.org/10.1348/135532509X474822

Brewer, N., & Wells, G. L. (2006). The confidence-accuracy relationship in eyewitness identification: Effects of lineup instructions, foil similarity, and target-absent base rates. *Journal of Experimental Psychology: Applied*, *12*, 11–30. http://doi.org/10.1037/1076-898X.12.1.11

Clark, S. E., & Godfrey, R. D. (2009). Eyewitness identification evidence and innocence risk. *Psychonomic Bulletin & Review*, *16*, 22–42. http://dx.doi.org/10.3758/PBR.16.1.22

Clark, S. E., Erickson, M. A., & Breneman, J. (2011). Probative value of absolute and relative judgments in eyewitness identification. *Law and Human Behavior*, *35*, 364–380. http://dx.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,* 1227–1239. http://doi.org/10.1177/0956797616655789

Colloff, M. F., Wade, K. A., Strange, D., & Wixted, J. T. (in press). Filler Siphoning Cannot Explain The Fair Lineup Advantage: Reply to Smith, Wells, Smalarz, and Lampinen (2018). *Psychological Science.*

Colloff, M. F., Wade, K. A., Wixted, J. T., & Maylor, E. A. (2017). A signal-detection analysis of eyewitness identification across the adult lifespan. *Psychology and Aging*, *32*, 243–258. http://doi.org/10.1037/pag0000168

Davis, J. P., Valentine, T., Memon, A., & Roberts, A. J. (2015). Identification on the street: A field comparison of police street identifications and video line-ups in England. *Psychology, Crime & Law*, *21*, 9–27. http://dx.doi.org/10.1080/1068316X.2014.915322

Duncan, M. (2006). *A signal detection model of compound decision tasks.* (Tech Note DRDC TR 2006-25). Toronto, ON: Defence Research and Development Canada.

Flowe, H. D., Klatt, T., & Colloff, M. F. (2014). Selecting fillers on emotional appearance improves lineup identification accuracy. *Law and Human Behavior*, *38*, 509–519. http://dx.doi.org/10.1037/lhb0000101

Gonzalez, R., Ellsworth, P. C., & Pembroke, M. (1993). Response biases in lineups and showups. *Journal of Personality and Social Psychology*, *64*, 525–537. http://dx.doi.org/10.1037/0022-3514.64.4.525

Gronlund, S. D., Carlson, C. A., Neuschatz, J. S., Goodsell, C. A., Wetmore, S. A., Wooten, A., & Graham, M. (2012). Showups versus lineups: An evaluation using ROC analysis. *Journal of Applied Research in Memory and Cognition*, *1*, 221–228. http://dx.doi.org/10.1016/j.jarmac.2012.09.003

Gronlund, S. D., Mickes, L., Wixted, J. T., & Clark, S. E. (2015). Conducting an eyewitness lineup: How the research got it wrong. In B. Ross (Ed.), *The psychology of learning and motivation* (Vol. 63, pp. 1–43). New York, NY: Academic Press.

Gronlund, S. D., Wixted, J. T., & Mickes, L. (2014). Evaluating eyewitness identification procedures using Receiver Operating Characteristic analysis. *Current Directions in Psychological Science*, *23*, 3–10. doi:10.1177/0963721413498891

Horry, R., Palmer, M. A., & Brewer, N. (2012). Backloading in the sequential lineup prevents within-lineup criterion shifts that undermine eyewitness identification performance. *Journal of Experimental Psychology: Applied*, *18*, 346–360. http://dx.doi.org/10.1037/a0029779

Key, K. N., Cash, D. K., Neuschatz, J. S., Price, J. L., Wetmore, S. A., & Gronlund, S. D. (2015). Age differences (or lack thereof) in discriminability for lineups and showups. *Psychology, Crime & Law*, *21*, 871–889. http://dx.doi.org/10.1080/1068316X.2015.1054387

Lampinen, J. M. (2016). ROC analyses in eyewitness identification research. *Journal of Applied Research in Memory and Cognition, 5*, 21–33. doi:10.1016/j.jarmac.2015.08.006

Lindsay, D. S., Read, J. D., & Sharma, K. (1998). Accuracy and confidence in person identification: The relationship is strong when witnessing conditions vary widely. *Psychological Science*, *9*, 215–218. http://dx.doi.org/10.1111/1467-9280.00041

Lindsay, R. C. L., Pozzulo, J. D., Craig, W., Lee, K., & Corber, S. (1997). Simultaneous lineups, sequential lineups, and showups: Eyewitness identification decisions of adults and children. *Law and Human Behavior*, *21*, 391–404. http://dx.doi.org/10.1023/A:1024807202926

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

Mickes, L., Moreland, M. B., Clark, S. E., & Wixted, J. T. (2014). Missing the information needed to perform ROC analysis? Then compute d′, not the diagnosticity ratio. *Journal of Applied Research in Memory and Cognition*, *3*, 58–62. http://doi.org/10.1016/j.jarmac.2014.04.007

Mickes, L., Flowe, H. D., & Wixted, J. T. (2012). Receiver Operating Characteristic Analysis of Eyewitness Memory: Comparing the Diagnostic Accuracy of

Simultaneous vs. Sequential Lineups. *Journal of Experimental Psychology: Applied, 18,* 361–376.

Neuschatz, J. S., Wetmore, S. A., Key, K., Cash, D., Gronlund, S. D., & Goodsell, C. A. (2016). A Comprehensive evaluation of showups. In M. Miller & B. Bornstein (Eds.), *Advances in Psychology and Law.* New York, NY: Springer.

Palmer, M. A., Brewer, N., & Weber, N. (2010). Postidentification feedback affects subsequent eyewitness identification performance. *Journal of Experimental Psychology: Applied, 16,* 387–398. http://dx.doi.org/10.1037/a0021034

Palmer, M. A., Brewer, N., Weber, N., & Nagesh, A. (2013). The confidence-accuracy relationship for eyewitness identification decisions: Effects of exposure duration, retention interval, and divided attention. *Journal of Experimental Psychology: Applied*, *19*, 55–71. http://dx.10.1037/a0031602

Police Executive Research Forum. (2013). *A national survey of eyewitness identification procedures in law enforcement agencies.* Washington, DC: Police Executive Research Forum. Retrieved from http://policeforum.org/library/eyewitness- identification/NIJEyewitnessReport.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,* 77–84. http://dx.doi.org/10.1186/1471-2105-12-77

Rotello, C. M., & Chen, T. (2016). ROC curve analyses of eyewitness identification decisions: An analysis of the recent debate. *Cognitive Research: Principles and Implications*, *1*, 10. http://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*, 944–954. http://doi.org/10.3758/s13423-014-0759-2

Sauer, J., Brewer, N., Zweck, T., & Weber, N. (2010). The effect of retention interval on the confidence-accuracy relationship for eyewitness identification. *Law and Human Behavior*, *34*, 337–347. http://doi.org/10.1007/s10979-009-9192-x

Smith, A. M., Lampinen, J. M., Wells, G. L., Smalarz, L., &. Mackovichova, S. (in press). Deviation from Perfect Performance Measures the Diagnostic Utility of Eyewitness Lineups but Partial Area Under the ROC Curve Does Not. *Journal of Applied Research in Memory and Cognition.*

Smith, A. M., Wells, G. L., Lindsay, R. C. L., & Penrod, S. D. (2017). Fair lineups are better than biased lineups and showups, but not because they increase underlying discriminability. *Law and Human Behavior, 41,* 127–145. http://doi.org/10. 1037/lhb0000219

Smith, A. M., Wells, G. L., Smalarz, L., & Lampinen, J. M. (2018). Increasing the Similarity of Lineup Fillers to the Suspect Improves the Applied Value of Lineups Without Improving Memory Performance: Commentary on Colloff, Wade, & Strange (2016). *Psychological Science.*

Sporer, S. L., Penrod, S., Read, D., & Cutler, B. (1995). Choosing, confidence, and accuracy: A meta-analysis of the confidence-accuracy relation in eyewitness identification studies. *Psychological Bulletin*, *118*, 315–327. http://dx.doi.org/10.1037/0033-2909.118.3.315

Steblay, 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*, 523–540. http://dx.doi.org/10.1023/A:1025438223608

Stovall v. Denno. (1967). 388 U.S. 293.

Valentine, T., & Heaton, P. (1999). An evaluation of the fairness of police line-ups and video identifications. *Applied Cognitive Psychology*, *13*, S59–S72. http://dx.doi.org/10.1002/(SICI)1099-0720(199911)13:1+<S59::AID-ACP679>3.0.CO;2-Y

Valentine, T., Davis, J. P., Memon, A., & Roberts, A. (2012). Live showups and their influence on a subsequent video line-up. *Applied Cognitive Psychology*, *26*, 1–23. http://dx.doi.org/10.1002/acp.1796

Wells, G. L. (2001). Eyewitness lineups: Data, theory, and policy. *Psychology, Public Policy, and Law, 7,* 791–801. http://dx.doi.org/10.1037/1076-8971.7.4.791

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 and Cognition*, *4*, 313–317. http://dx.doi.org/10.1016/j.jarmac.2015.08.008

Wells, G. L., Smith, A. M., & Smalarz, L. (2015). ROC analysis of lineups obscures information that is critical for both theoretical understanding and applied purposes. *Journal of Applied Research in Memory and Cognition*, *4*, 324–328. http://dx.doi.org/10.1016/j.jarmac.2015.08.010

Wetmore, S. A., Neuschatz, J. S., Gronlund, S. D., Wooten, A., Goodsell, C. A., & Carlson, C. A. (2015). Effect of retention interval on showup and lineup performance. *Journal of Applied Research in Memory and Cognition*, *4*, 8–14. http://dx.doi.org/10.1016/j.jarmac.2014.07.003

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. http://dx.doi.org/10.1186/s41235-017-0084-1

Whitney, D. & Leib, A. Y. (2018). Ensemble Perception. *Annual Review of Psychology, 69,* 105-129. https://doi.org/10.1146/annurev-psych-010416-044232

Wilson, B. M., Donnelly, K., Christenfield, N., & Wixted, J.T. (under review). *Making Sense of Sequential Lineups: An Experimental and Theoretical Analysis of Position Effects.*

Wixted, J. T., & Mickes, L. (2014). A signal-detection-based diagnostic-feature-detection model of eyewitness identification. *Psychological Review*, *121*, 262–276. http://dx.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 and Cognition*, *4*, 329–334. http://dx.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*, 3:9. http://doi.org/10.1186/s41235-018-0093-8

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

Wolchover, D., & Heaton-Armstrong, A. (2014). Street Identification. *Criminal Law & Justice Weekly, 178*(10)*,* retrieved from http://www.criminallawandjustice.co.uk/features/Street-Identification

Zarkadi, T., Wade, K. A., & Stewart, N. (2009). Creating fair lineups for suspects with distinctive features. *Psychological Science*, *20*, 1448–1453. http://doi.org/10.1111/j.1467-9280.2009.02463.x

Appendix A

Experiment 2: Signal-detection Model

Although we had originally planned to fit a model to the chooser data (i.e., Yes responses), here we describe a model that we fit to all of the empirical data—choosers and non-choosers—to improve our theoretical understanding about non-choosers. Our pre-planned model fits are available in the online supplemental materials. Notably, with respect to choosers—which is what the diagnostic-feature-detection and filler siphoning theories make a prediction about—both model-fitting exercises lead to the same conclusions.

The chooser and non-chooser model assumes that the memory strength values for innocent suspects and guilty suspects have Gaussian distributions with means of $\mu_{innocent}$ and μ_{guilty}, respectively. The distance between the $\mu_{innocent}$ and μ_{guilty} distributions reflects underlying theoretical discriminability, with a greater overlap of distributions reflecting a poorer ability to discriminate between innocent and guilty suspects. The model also assumes that there is a set of response criteria that reflect different levels of confidence. To limit the number of parameters, we collapsed our data to a 5-point confidence scale. We combined confidence ratings of No 80-70 (c_1), No 60-0 (c_2), Yes 0-60 (c_3), Yes 70-80 (c_4), and Yes 90-100 (c_5). Once these five categories were specified, the number of non-choosers who gave a confidence rating of 90-100 was fixed. The model assumes that a positive identification is made when the suspect's face is familiar enough to exceed c_3, and the confidence in the identification is determined by the highest criterion that is exceeded. An illustration of the model is presented in Figure A1. Target-present showups each had 5 degrees of freedom because there were 5 levels of confidence for guilty suspect decisions. Target-absent showups each had 5 degrees of freedom because there were 5 levels of confidence for innocent suspect decisions. Thus, for both simultaneous showups and standard showups there were 5 + 5 = 10 degrees of freedom in the data. To fit the model, we minimized the chi-square goodness-of-fit statistic, and we fixed $\mu_{innocent}$ and $\sigma_{innocent}$ to 0 and 1, respectively.

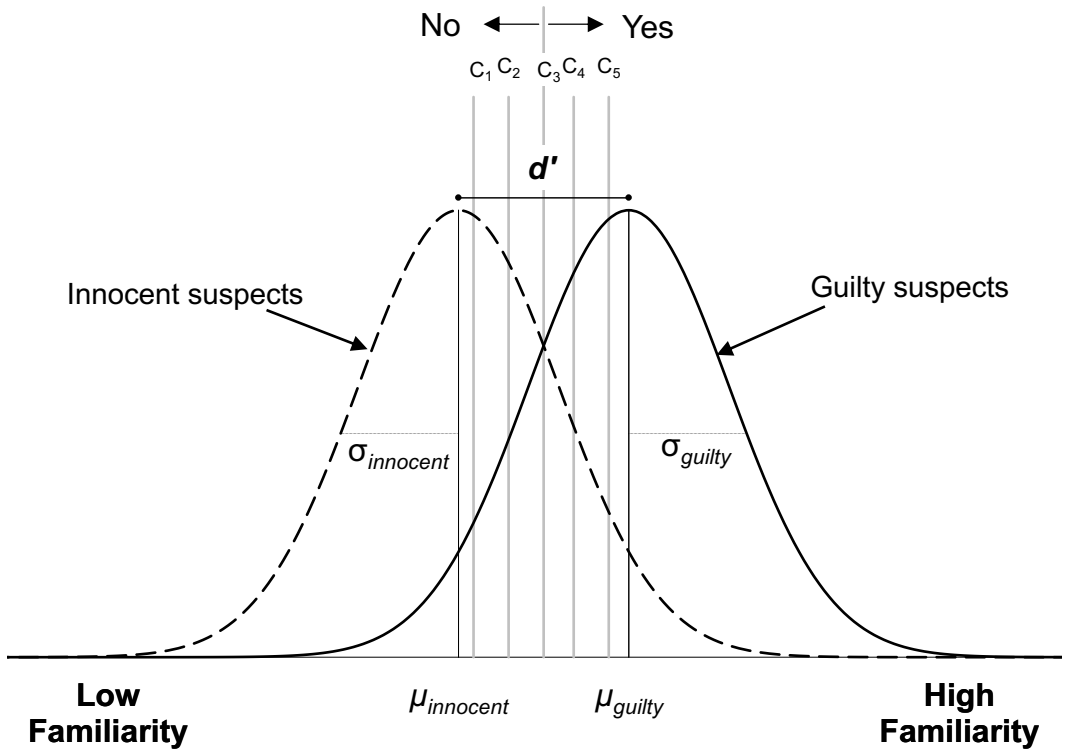

Figure A1. Equal-variance signal-detection model for a showup when fitting both chooser and non-chooser data. The distance between the $\mu_{innocent}$ and μ_{guilty} distributions measures underlying theoretical discriminability, measured here by *d'*.

To begin, we fit a full unequal-variance model which allowed the mean and standard deviation of the guilty suspect distribution to differ across choosers and non-choosers, as well as across simultaneous showups and standard showups. This model had 18 parameters ($\mu_{guilty(choosers)}$, $\mu_{guilty(non-choosers)}$, $\sigma_{guilty(choosers)}$, $\sigma_{guilty(non-choosers)}$, c_1, c_2, c_3, c_4, c_5 for both showup conditions) and there were 20 degrees of freedom in the data (10 degrees of freedom in both showup conditions). As such, the fit of the full model to the data involved 20 – 18 = 2 degrees of freedom. The model fit statistics are presented in the *full model* column in Table A1. We then took this full model and determined how many free parameters could be eliminated without making the fit significantly worse.

To search for the simplest reduced model that did not fit significantly worse, we first constrained μ_{guilty} (i.e., *d*) and σ_{guilty} in each showup condition to be the same in choosers and non-choosers. The model had 14 parameters ($\mu_{guilty(choosers\ \&\ non-choosers)}$, $\sigma_{guilty(choosers\&\ non-choosers)}$, c_1, c_2, c_3, c_4, c_5 for both showup conditions), so the fit of the model to the data involved 20 – 14 = 6 degrees of freedom. The model fit statistics for the *constrained choosers and non-choosers model* are presented in Table A1. The constrained choosers and non-choosers model provided a significantly worse fit of the data than the full model, χ^2 (4) = 32.03, $p < .001$. This indicates that choosers and non-choosers

are drawn from different Gaussian distributions—they differ significantly in terms of discriminability (i.e., d and σ_{guilty})—and, as such, should be separately analyzed.

For choosers, the simplest reduced model that did not fit significantly worse was an unequal variance model, with σ_{guilty} and μ_{guilty} allowed to vary across the showup conditions. For non-choosers, the simplest reduced model was an equal variance model (i.e., $\sigma_{guilty} = \sigma_{innocent} = 1$), with μ_{guilty} (i.e., d') constrained across the showup conditions. The overall model had 15 parameters ($\mu_{guilty(choosers\_simultaneous\_showup)}$, $\mu_{guilty(choosers\_standard\_showup)}$, $\sigma_{guilty(choosers\_simultaneous\_showup)}$, $\sigma_{guilty(choosers\_standard\_showup)}$, $\mu_{guilty(non\text{-}choosers\_simultaneous\_\&\_standard\_showup)}$, and c_1, c_2, c_3, c_4, c_5 for both showup conditions), so the fit of the model to the data involved 20 – 15 = 5 degrees of freedom. The model fit statistics presented in the *best fitting reduced model* column in Table A1 show that the model fit the data well. We plotted the lines on Figure 2A using the *best fitting reduced model* predicted values. The model-predicted lines run through the empirical data points, which provides further evidence that the model was able to account for the trends in our data.

What does the best-fitting reduced model tell us about the effect of presenting similar looking faces around the suspect? In choosers, the best-fitting reduced model estimated d and σ to be higher in the simultaneous showup than in the standard showup (see best-fitting reduced model column in Table A1). In each showup condition, we used the model-predicted d and σ values to calculate d_a, which is the relevant discriminability measure when the magnitude of the unequal variance parameter differs across conditions, using the formula $d_a = (\mu_{guilty} - \mu_{innocent}) / \sqrt{[.5(\sigma_{guilty}^2 + \sigma_{innocent}^2)]}$. Setting $\mu_{innocent} = 0$ and $\sigma_{innocent} = 1$ by convention, the equation reduces to $d_a = \mu_{guilty} / \sqrt{[.5(\sigma_{guilty}^2 + 1)]}$. As predicted, d_a was larger in the simultaneous showup ($d_a = 1.54$) than in the standard showup ($d_a = 1.45$). The difference in d_a across the showup conditions is statistically significant, because when we constrained μ_{guilty} and σ_{guilty} (i.e., d_a) to be equal across the simultaneous and standard showups in choosers (see constrained choosers μ and σ column in Table A1), this provided a significantly worse fit of the data than the best-fitting reduced model, χ^2 (2) = 6.41, p = .04. This indicates that, in choosers, presenting similar-looking faces alongside the suspect enhances people's ability to discriminate between innocent and guilty suspects. Yet, in non-choosers, a different story emerged. When we constrained μ_{guilty} and σ_{guilty} (i.e., d') to be the same across the simultaneous and standard showups in the best-fitting reduced model, this did not significantly worsen the fit compared to the full model in which μ_{guilty} and σ_{guilty} were free to vary across the simultaneous and standard showups, χ^2 (3) = 3.16, p = .37. This indicates that, in non-choosers, d' was not significantly different across the showup conditions. That is, in non-choosers, presenting similar-looking faces alongside the suspect did not enhance people's ability to discriminate between innocent and guilty suspects.

Table A1

Model Fits for the Simultaneous Showup vs. Standard Showup Comparisons (Experiment 2)

	Full model		Constrained choosers and non-choosers		Best-fitting reduced model		Constrained choosers' μ and σ	
Estimate	Simultaneous Showup	Standard Showup	Simultaneous Showup	Standard Showup	Simultaneous Showup	Standard Showup	Simultaneous Showup	Standard Showup
Choosers								
μ_{guilty}	1.46	1.19	1.45	1.12	1.43	1.20	1.28	1.28
σ_{guilty}	0.83	0.60	1.15	0.93	0.85	0.61	0.68	0.68
Non-choosers								
μ_{guilty}	1.54	1.35	1.45	1.12	1.21	1.21	1.20	1.20
σ_{guilty}	1.25	1.14	1.15	0.93	1.00	1.00	1.00	1.00
c_1	–0.49	–0.53	–0.48	–0.51	–0.48	–0.51	–0.49	–0.50
c_2	0.24	0.11	0.23	0.09	0.21	0.12	0.19	0.12
c_3	1.09	0.84	1.01	0.70	1.08	0.85	1.00	0.89
c_4	1.45	1.22	1.46	1.19	1.44	1.22	1.30	1.30
c_5	1.99	1.67	2.16	1.83	1.99	1.68	1.74	1.80
Overall χ^2	2.52		34.55		5.68		12.10	
Overall df	2		6		5		7	
Overall *p*	.28		<.001		.34		.10	

Note. The full model allows μ_{guilty} and σ_{guilty} to differ across choosers and non-choosers, as well as across simultaneous and standard showups. The constrained choosers and non-choosers model equates both μ_{guilty} and σ_{guilty} in choosers and non-choosers. The best-fitting reduced model allows μ_{guilty} and σ_{guilty} to differ across the simultaneous and standard showup in choosers, but sets σ_{guilty} to 1 and equates μ_{guilty} across the simultaneous and standard showups in non-choosers. The constrained choosers' μ and σ model is the same as the best-fitting reduced model, but equates μ_{guilty} and σ_{guilty} (i.e., equates the discriminability parameter d_a) across the simultaneous and standard showups in choosers. The overall χ^2, df and *p* rows are the goodness-of-fit statistics for each model.

Appendix B

Experiment 3: Signal-detection Model

Although we also had a standard simultaneous lineup condition in Experiment 3, in essence, we fit the same model that we used to fit the showup data in Experiment 2. The model for a fair lineup—like the model for a showup—assumes that the memory strength values are represented by two Gaussian distributions. In a fair lineup, the innocent suspect and fillers are equally similar to the guilty suspect. Thus, the innocent suspect and the fillers are both drawn from the same memory strength distribution, with a mean of $\mu_{innocent}$. The guilty suspect is drawn from a different memory strength distribution, with a mean of μ_{guilty}. As before, we measured the distance between the $\mu_{innocent}$ and μ_{guilty} distributions. When applied to the fair lineup case, this measure can be thought to reflect both subjects' ability to discriminate (a) guilty suspects from innocent suspects and (b) guilty suspects from innocent suspects *and* fillers. This is because, in a fair lineup, innocent suspects and fillers have the same memory strength value, on average.

Again, the model assumes that there are a set of response criterion that reflect different levels of confidence. For a showup, the model assumes that a positive identification is made when the suspect's face is familiar enough to exceed the lowest decision criterion (c_1). For a lineup, the model assumes that a positive identification is made when the familiarity value of the most familiar face in the lineup exceeds the lowest decision criterion (c_1). This has been called the independent-observation rule (Macmillan & Creelman, 2005), or the best-above-criterion rule (e.g., Clark, Erickson, & Breneman, 2011; Colloff, Wade, Strange, & Wixted, 2018) elsewhere in the lineup literature. Another decision rule that has been used in the lineup literature is the *integration* rule. The integration rule assumes that the most familiar face is positively identified when the sum of the familiarity values of all of the faces in the lineup exceeds the lowest decision criterion (e.g., Duncan, 2006; Palmer, Brewer, & Weber, 2010). To date, only a small amount of research has compared the efficacy of the independent-observation and integration decision rules, but that research suggests that the independent-observation decision rule may better account for the decision-making process of eyewitnesses (Colloff, Wade, Strange, & Wixted, 2018; Wixted, Vul, Mickes, & Wilson, 2018). It is also a simple decision rule that can be applied to both showups and lineups. As such, here, we used the independent-observation decision rule.[3] In both the showup and lineup case, the confidence in the identification is determined by the highest criterion that is exceeded. To limit the number of parameters, we collapsed our data to a 5-point confidence scale. We combined confidence ratings of

[3] It is important to highlight that the lineup model that we used is a compound signal-detection model as defined by Duncan (2006), because it has both detection and identification components. That is, it assumes that people first detect the most familiar lineup member and then identify that individual if that face exceeds the lowest decision criterion.

Yes 0-20 (c_1), Yes 30-40 (c_2), Yes 50-60 (c_3), Yes 70-80 (c_4), and Yes 90-100 (c_5). Once these five categories are specified, the number of non-choosers (i.e., subjects who said "No the culprit is not here") was fixed.

We fit an unequal variance model to our data. Both the simultaneous showup and standard showup models had 7 parameters (μ_{guilty}, σ_{guilty}, c_1, c_2, c_3, c_4, c_5) and, in each, there were 10 degrees of freedom in the data (5 confidence levels for guilty suspect identifications and 5 degrees of freedom for innocent suspect identifications). As such, the fit of the *full model* to each of the showup conditions had 10 – 7 = 3 degrees of freedom. The lineup model also had 7 parameters (μ_{guilty}, σ_{guilty}, c_1, c_2, c_3, c_4, c_5) and there were 15 degrees of freedom in the data (5 confidence levels for guilty suspect identifications and 5 confidence levels for filler identifications from target-present lineups and 5 confidence levels for innocent suspect or filler identifications from target-absent lineups). As such, the fit of the full unequal-variance model to the lineup condition had 15 – 7 = 8 degrees of freedom.

To fit the model to each condition, we minimized the chi-square goodness-of-fit statistic and we fixed $\mu_{innocent}$ and $\sigma_{innocent}$ to 0 and 1, respectively. The chi-square goodness-of-fit statistics showed that the unequal-variance model fit the simultaneous showup, $\chi^2(3) = 2.28$, $p = .52$, and standard showup data, $\chi^2(3) = 4.55$, $p = .21$, but the predictions of the model departed significantly from the observed standard simultaneous lineup data, $\chi^2(8) = 42.73$, $p<.001$. Nevertheless, we used the model predicted values to draw lines of best fit on the *p*ROC plot (Figure 4A) and it is clear from the correspondence between the model-predicted line of best fit and the observed data, that the model was able to capture the basic trends in our data, even in the standard lineup condition.

Table A2 shows the best-fitting parameters and the chi-square goodness-of-fit statistics for the full unequal-variance model (full model column). For each condition, we used the model predicted μ_{guilty} (i.e., d) and σ values to calculate d_a, which is the relevant discriminability measure when the magnitude of the unequal variance parameter differs across conditions, again using the formula $d_a = (\mu_{guilty} - \mu_{innocent}) / \sqrt{[.5(\sigma_{guilty}^2 + \sigma_{innocent}^2)]}$. As predicted by the diagnostic-feature-detection theory, d_a was larger in the simultaneous showup ($d_a = 1.56$) and the standard simultaneous lineup ($d_a = 1.25$) than in the standard showup ($d_a = 1.23$). To test whether the differences in d_a were statistically significant, we made three pairwise comparisons in which we constrained μ_{guilty} and σ_{guilty} (i.e., d_a) to be equal across the two conditions that were being compared (while allowing the confidence criteria to differ across conditions). Table A2 shows the best-fitting parameters and the chi-square goodness-of-fit statistics for the constrained model (constrained model column). In comparison to the full model, the constrained model provided a significantly worse fit to the simultaneous showup and standard showup data, $\chi^2 (2) = 8.94$, $p = .01$, and the lineup and standard showup data, $\chi^2 (2) = 18.43$,

$p < .001$. This indicates that presenting similar-looking faces alongside the suspect enhances people's ability to discriminate between innocent and guilty suspects. Interestingly, the constrained model provided a significantly worse fit to the simultaneous showup and standard simultaneous lineup data, χ^2 (2) = 7.30, $p = .03$, indicating that ability to discriminate between innocent and guilty suspects was estimated to be better in the simultaneous showup than the standard simultaneous lineup. Taken together, the results of the model-fitting exercise are broadly concordant with the results of our analyses based on the atheoretical *p*AUC measure. Most importantly, we replicated the key result—underlying theoretical discriminability was better in the simultaneous showup than the standard showup, despite there being no opportunity for erroneous identifications to be spread across the fillers, as predicted by the diagnostic-feature-detection account.

Table A2

Full and Constrained Model fits for the Simultaneous Showup, Standard Showup and Standard Simultaneous Lineup Comparisons (Experiment 3)

	Full Model		Constrained Model	
Estimate	Simultaneous Showup	Standard Showup	Simultaneous Showup	Standard Showup
μ_{guilty} (d)	1.35	1.05	1.18	1.18
σ_{guilty}	0.71	0.68	0.71	0.71
c_1	1.09	0.42	0.95	0.49
c_2	1.11	0.44	0.98	0.51
c_3	1.17	0.52	1.03	0.59
c_4	1.42	0.81	1.28	0.89
c_5	1.86	1.37	1.71	1.48
Overall χ^2	6.83		15.76	
Overall df	6		8	
Overall p	.34		.05	

	Full Model		Constrained Model	
	Simultaneous Showup	Standard Lineup	Simultaneous Showup	Standard Lineup
μ_{guilty} (d)	1.35	1.27	1.32	1.32
σ_{guilty}	0.71	1.03	0.94	0.94
c_1	1.09	1.03	1.01	1.04
c_2	1.11	1.13	1.05	1.14
c_3	1.17	1.33	1.11	1.34
c_4	1.42	1.73	1.42	1.73
c_5	1.86	2.35	1.97	2.32
Overall χ^2	45.01		52.31	
Overall df	11		13	
Overall p	<.001		<.001	

	Full Model		Constrained Model	
	Standard Showup	Standard Lineup	Standard Showup	Standard Lineup
μ_{guilty} (d)	1.05	1.27	1.23	1.23
σ_{guilty}	0.68	1.03	0.73	0.73
c_1	0.42	1.03	0.33	1.06
c_2	0.44	1.13	0.40	1.13
c_3	0.52	1.33	0.57	1.30
c_4	0.81	1.73	0.94	1.67
c_5	1.37	2.35	1.58	2.31
Overall χ^2	47.28		65.71	
Overall df	11		13	
Overall p	<.001		<.001	

Note. The full model allows the discriminability parameters (μ_{guilty} and σ_{guilty}) to differ between the two conditions being compared. The constrained model holds the discriminability parameters constant across the two conditions being compared. Overall χ^2, df and p rows represent goodness-of-fit statistics when the model was fit both conditions.