

# UC San Diego

## UC San Diego Previously Published Works

### Title

The Reveal Procedure: A Way to Enhance Evidence of Innocence From Police Lineups

### Permalink

<https://escholarship.org/uc/item/5j8597h3>

### Journal

Law and Human Behavior, 46(2)

### ISSN

0147-7307

### Authors

Yilmaz, Anne S  
Lebensfeld, Taylor C  
Wilson, Brent M

### Publication Date

2022-04-01

### DOI

10.1037/lhb0000478

### Copyright Information

This work is made available under the terms of a Creative Commons Attribution-NonCommercial-NoDerivatives License, available at <https://creativecommons.org/licenses/by-nc-nd/4.0/>

Peer reviewed

# The Reveal Procedure: A Way to Enhance Evidence of Innocence From Police Lineups

Anne S. Yilmaz, Taylor C. Lebensfeld, and Brent M. Wilson  
Department of Psychology, University of California, San Diego



**Objective:** Recent work has established that high-confidence identifications (IDs) from a police lineup can provide compelling evidence of guilt. By contrast, when a witness rejects the lineup, it may offer only limited evidence of innocence. Moreover, confidence in a lineup rejection often provides little additional information beyond the rejection itself. Thus, although lineups are useful for incriminating the guilty, they are less useful for clearing the innocent of suspicion. Here, we test predictions from a signal-detection-based model of eyewitness ID to create a lineup that is capable of increasing information about innocence. **Hypotheses:** Our model-based simulations suggest that high-confidence rejections should exonerate many more innocent suspects and do so with higher accuracy if, after a witness rejects a lineup but before they report their confidence, they are shown the suspect and asked, “How sure are you that this person is *not* the perpetrator?” **Method:** Participants ( $N = 3,346$ ) recruited from Amazon Mechanical Turk watched a 30-s mock-crime video of a perpetrator. Afterward, they were randomly assigned to lineup procedures using a 2 (standard control vs. reveal condition)  $\times$  2 (target present vs. target absent) design. A standard simultaneous lineup served as the control condition. The reveal condition was identical to the control condition except in cases of lineup rejection: When a lineup rejection occurred, the suspect appeared on the screen, and participants provided a confidence rating indicating their belief that the suspect was not the perpetrator. **Results:** The reveal procedure increased both the accuracy and frequency of high-confidence rejections relative to the standard simultaneous lineup. **Conclusions:** Collecting a confidence rating about the suspect after a lineup is rejected may make it possible to quickly clear innocent suspects of suspicion and reduce the amount of contact that innocent people have with the legal system.

### Public Significance Statement

We found that changing the standard lineup procedure may allow a greater number of innocent suspects to quickly be cleared of suspicion. The procedural change, which is easily implemented, is simply this: When a lineup is rejected, but before the witness is asked about their confidence, the suspect is revealed to them along with this question: “How sure are you that this person is *not* the perpetrator?”

**Keywords:** confidence–accuracy, eyewitness confidence, eyewitness identification, lineup rejections, signal detection theory

**Supplemental materials:** <https://doi.org/10.1037/lhb0000478.supp>

Scientists have long endeavored to improve photo lineup procedures, such as by testing whether instructions should be biased or unbiased (e.g., Cutler et al., 1987) or whether photos should be presented simultaneously or sequentially (e.g., Lindsay & Wells, 1985). To

investigate such issues, recent research has relied on a measurement methodology grounded in signal detection theory known as receiver operating characteristic (ROC) analysis (Green & Swets, 1966; Mickes et al., 2012). Although the details of exactly how to

Lora Levett served as Action Editor.

This article was published Online First January 27, 2022.


Anne S. Yilmaz  <https://orcid.org/0000-0003-4759-4910>


Taylor C. Lebensfeld  <https://orcid.org/0000-0002-3630-7886>

Brent M. Wilson  <https://orcid.org/0000-0001-9018-7902>

The authors have no conflicts of interest to disclose. This study was supported by a grant from Arnold Ventures. The funding source had no role other than financial support. Some of these data were presented at Psychological Society’s 61st Annual Convention (2020, virtual) and the Association

for Psychological Science’s Annual Convention (2021, virtual).

 The data are available at <https://osf.io/vwza7>.

 The preregistered design and analysis plan is accessible at <https://aspredicted.org/c6nh4.pdf>.

Correspondence concerning this article should be addressed to Brent M. Wilson, Department of Psychology, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093, United States. Email: [b6wilson@ucsd.edu](mailto:b6wilson@ucsd.edu)

perform ROC analysis remain debated (e.g., Smith et al., 2019, 2020), it is safe to say that signal detection theory is useful for measurement purposes (Lee & Penrod, 2019).

In addition to providing a useful measurement methodology, signal detection theory can be used to generate novel predictions about how to improve lineup procedures. We follow that approach here. Our specific goal is to increase the information value of lineup rejections, which occur when a witness declares that no one in the lineup is the perpetrator. Prior research on this issue suggests that lineup rejections provide modest evidence of innocence (Wells & Olson, 2002; Wells et al., 2015). However, although confidence in a positive identification (ID) is highly indicative of accuracy (Wixted & Wells, 2017), confidence in a lineup rejection often provides only limited additional information beyond the rejection itself (e.g., Brewer & Wells, 2006). Thus, unlike other forms of forensic evidence (e.g., fingerprints), lineups are not particularly useful for removing suspicion from the innocent. Why that is remains unclear.

One reason that high-confidence lineup rejections may not provide strong evidence of innocence may be related to the number of faces to which the confidence rating is applied. Specifically, whereas confidence for a positive ID is made in relation to a single face in the lineup, confidence for a lineup rejection may be made in relation to the entire set of faces in a lineup. In agreement with this idea, Sauerland et al. (2012) observed a strong confidence–accuracy relationship for rejections in a showup procedure (i.e., wherein only one suspect photo is presented without accompanying filler photos). Additionally, Lindsay et al. (2013) found that confidence ratings are less informative when made to a set of target faces as opposed to a single target face.

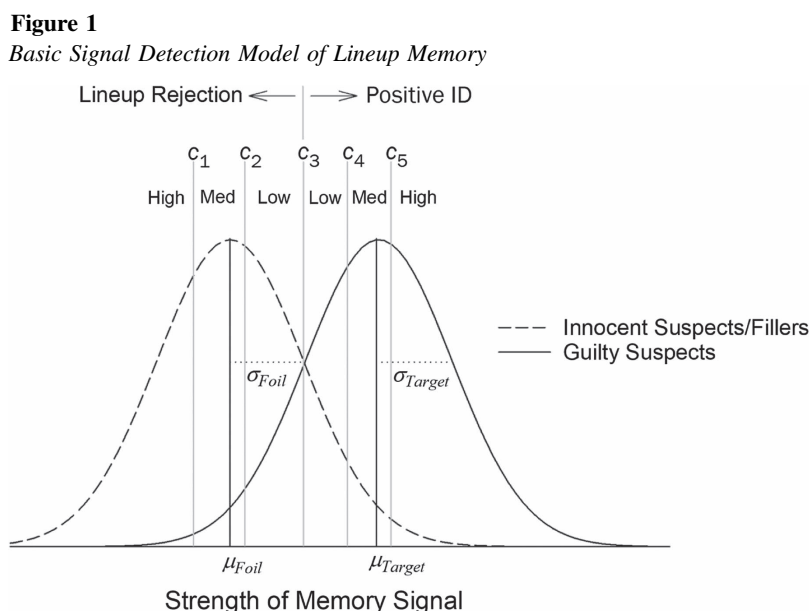
To shed light on why confidence in decisions about a single face is more diagnostic of accuracy than confidence in decisions about a set

of faces, we made use of signal detection theory. Figure 1 illustrates the standard signal detection model of a proper photo lineup. A proper photo lineup consists of one suspect (the person the police believe may have committed the crime) and five or more physically similar fillers who, at a minimum, match the description of the perpetrator provided by the eyewitness (Wells et al., 2020). The fillers are known to be innocent, and the purpose of the lineup test is to help the police determine whether the suspect is innocent or guilty.

For the model shown in Figure 1, a six-person target-present (TP) lineup is conceptualized as one random draw from the target distribution (the memory signal associated with the guilty suspect) and five random draws from the foil distribution (the memory signals associated with the five fillers). A six-person target-absent (TA) lineup is conceptualized as six random draws from the foil distribution. If the face that generates the strongest signal (the MAX face, whether suspect or filler) is strong enough to exceed the decision criterion ( $c_3$ ), then an ID of that face will be made. If none of the faces in the lineup generates a memory signal strong enough to exceed the decision criterion, then the lineup is rejected.

For positive IDs, additional decision criteria are used to rate the level of confidence associated with the MAX face. If that memory signal exceeds the rightmost confidence criterion in Figure 1 ( $c_5$ ), then a positive ID is made with high confidence. If the memory signal associated with the MAX face does not exceed  $c_5$  but does exceed  $c_4$ , the positive ID is made with medium confidence. Finally, if the memory signal does not exceed  $c_4$  but does exceed  $c_3$ —the decision criterion for making an ID or not—then the positive ID is made with low confidence.

Note that a positive ID and its corresponding confidence rating are made in relation to only one face in the lineup, namely the MAX face



*Note.* The mean of the target distribution ( $\mu_{target}$ ) is higher than the mean of the foil distribution ( $\mu_{foil}$ ) because, on average, the guilty suspect (i.e., the target) matches the representation of the perpetrator in memory better than the fillers and innocent suspect (i.e., the foils) do. An equal-variance model is assumed for simplicity ( $\sigma_{target} = \sigma_{foil}$ ), and there are five confidence criteria ( $c_1$  through  $c_5$ ), with  $c_3$  being the criterion that separates positive identifications (IDs) from lineup rejections.

(assumed here to be in accordance with the independent observations model; Wixted et al., 2018). By contrast, if the memory signal generated by the MAX face does not exceed the decision criterion, in which case the lineup as a whole is rejected, then the confidence rating is not necessarily applied to a single face and may instead be applied to the entire collection of faces in the lineup. How do witnesses determine confidence when all faces in the lineup are simultaneously rejected?

One possibility is that, when rejecting the lineup, witnesses base their confidence on the average memory signal generated by the faces in the lineup (Lindsay et al., 2013). According to this idea, if the average memory signal generated by the faces in the lineup is so weak that it falls below the leftmost confidence criterion ( $c_1$ ) in Figure 1, then the lineup is rejected with high confidence. If it falls above  $c_1$  but below  $c_2$ , the lineup is rejected with medium confidence. Finally, if it falls above  $c_2$  but below  $c_3$ , the lineup is rejected with low confidence.

This account of confidence in lineup rejections is, of course, not the only possibility. For example, despite not having to identify any face when rejecting a lineup, witnesses could still implicitly identify the MAX face and then base their confidence rating on that memory signal alone. As far as we know, previous research does not support one theoretical conceptualization of confidence in lineup rejections over another. However, it seems reasonable to hypothesize that there might be something qualitatively different about positive IDs versus lineup rejections in that confidence in a positive ID is highly indicative of accuracy, whereas confidence in a lineup rejection is often much less so. Therefore, we adopt the working hypothesis that the qualitative difference in behavior reflects a qualitative difference in the nature of the underlying memory signal upon which confidence is based (i.e., a qualitative difference in the decision variable). Although this line of thinking informed the present research, precisely determining the decision variable is outside the scope of the present work.

In a pilot study involving simulated data, we used this signal detection model to determine if it could predict the asymmetrical confidence–accuracy relationship for positive IDs versus lineup rejections and then asked, under those conditions, what it predicted should happen if the suspect were revealed before the eyewitness expressed their confidence in a lineup rejection (which we label the *reveal procedure*). In that case, confidence in the lineup rejection would be based on the memory signal generated by that face alone. We tested those predictions empirically in three experiments.

### Simulation-Based Pilot Study

To simulate data from a signal detection model (Figure 1), we first chose specific parameter values for  $\mu_{\text{target}}$  and five confidence criteria (with  $\mu_{\text{foil}} = 0$  and  $\sigma_{\text{target}} = \sigma_{\text{foil}} = 1$ ) and then conducted a simulation study consisting of 50,000 six-person TP trials and 50,000 six-person TA trials. We used a large number of trials simply because it resulted in stable simulated results (i.e., the number of trials was large enough that the simulated results were very similar every time we ran the simulation). Each TP trial involved one random draw from the target distribution (the memory signal associated with the guilty suspect) and five random draws from the foil distribution (the memory signals associated with the five fillers). A six-person TA lineup involved six random draws from the

foil distribution, with one randomly designated as the memory signal of the innocent suspect.

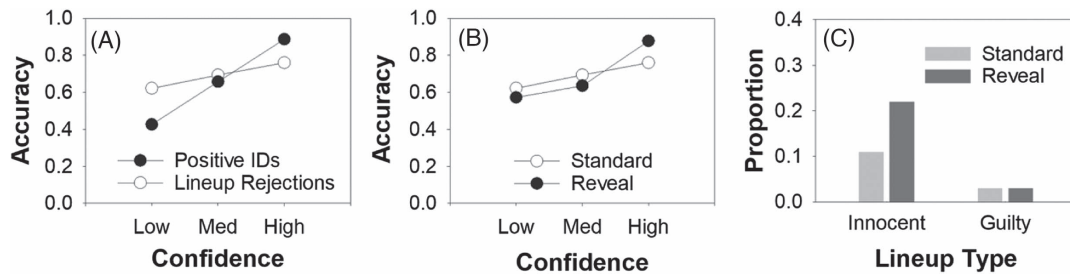
For both TP and TA lineups, on each trial, the maximum value of the six signals was determined first (the MAX signal), and the decision to make an ID or not was based on its magnitude. This model is technically known as the independent observations model (Wixted et al., 2018), and we used it because it is the simplest of various alternative signal detection models. According to this signal detection model, if the MAX signal exceeded  $c_3$ , a positive ID was made, with confidence determined by the highest criterion exceeded. For example, if the MAX signal was strong enough to exceed  $c_5$ , the positive ID was made with high confidence.

If the MAX signal did not exceed  $c_3$ , the lineup was rejected. What happened next differed between the standard and reveal conditions. In the standard condition, confidence in the lineup rejection was based on the average of the six memory signals. If the average memory signal generated by the faces in the lineup fell to the left of  $c_1$ , then the lineup was rejected with high confidence. If the average memory signal fell above  $c_1$  but below  $c_2$ , the lineup was rejected with medium confidence. Finally, if the average memory signal fell above  $c_2$  but below  $c_3$ , the lineup was rejected with low confidence. In the reveal condition, by contrast, confidence in the lineup rejection was based on the memory signal associated with the suspect (innocent or guilty).

We first searched for parameter settings (i.e.,  $\mu_{\text{target}}$  and  $c_1$  through  $c_5$ ) that generated an asymmetrical confidence–accuracy relationship of the kind often observed in the empirical literature. For this purpose, we computed calibration accuracy scores as they are typically computed. However, as described in more detail later, we collapsed the 0–100 confidence scale into low-, medium-, and high-confidence bins. Specifically, for positive IDs, accuracy is equal to the number of suspect IDs from TP lineups divided by the number of suspect IDs from TP lineups plus the number of filler IDs from TA lineups. This accuracy score differs from the accuracy score for confidence–accuracy characteristic analysis (Mickes, 2015), which is the hit rate (number of TP suspect IDs divided by the number of TP lineups) divided by the sum of the hit rate and the false alarm rate (number of TA suspect IDs or estimated TA suspect IDs divided by the number of TA lineups). Estimated suspect IDs are equal to the number of TA filler IDs divided by lineup size. We instead report calibration scores here because that was the measure used in studies our work is based on (e.g., Brewer & Wells, 2006). Similarly, for lineup rejections, accuracy is equal to the number of TA lineup rejections divided by the number of TA lineup rejections plus the number of TP lineup rejections. In all cases, the accuracy values were computed separately for each level of confidence.

We found a variety of parameter settings that could yield an asymmetric confidence–accuracy relationship similar to that often observed in empirical research. This outcome simply confirms that the model can yield that outcome, not that it necessarily predicts it. To illustrate its ability to generate asymmetrical confidence–accuracy relations for positive IDs versus lineup rejections, we can use the following parameter settings:  $\mu_{\text{target}} = 1.5$  and  $c_1$  through  $c_5 = -0.5, -0.3, 1.5, 2.0,$  and  $3.0$ , respectively. The simulation results based on these parameter settings for the standard lineup condition are depicted in Figure 2A. The confidence–accuracy function for lineup rejections (open circles) is not completely flat, but it is much flatter than the corresponding function for positive IDs.

**Figure 2**  
*Simulated Results From a Signal Detection Model of Lineups*



*Note.* Panel A: Confidence–accuracy relationship using the calibration method for positive IDs and lineup rejections (e.g., Brewer & Wells, 2006). Panel B: Confidence–accuracy relationship for lineup rejections from Panel A plotted with predicted confidence–accuracy relationship for each condition. Panel C: Proportion of innocent (target-absent) and guilty (target-present) target-present and target-absent lineups associated with high-confidence lineup rejections.

With the parameters still fixed at these values, we next asked what the model predicted about the accuracy of high-confidence rejections and the number of innocent suspects receiving a high-confidence rejection in the reveal condition. The results indicated that the accuracy of a lineup rejection made with high confidence should be higher in the reveal condition than in the standard condition (Figure 2B). In this particular simulation, accuracy increased from 76% correct (standard condition) to 89% correct (reveal condition). In addition, the number of innocent suspects who received a high-confidence lineup rejection should increase in the reveal condition compared to the standard condition (Figure 2C). In this particular simulation, the proportion of innocent suspects receiving a high-confidence lineup rejection increased from 11% in the standard condition to 22% in the reveal condition. The number of guilty suspects receiving high-confidence lineup rejections was not predicted to increase (3% for both the standard and the reveal conditions).

Because we did not test every conceivable parameter setting, we cannot claim that this model *necessarily* predicted these effects for the reveal condition. However, it did so in every scenario we investigated in which the parameter settings yielded an asymmetrical confidence–accuracy pattern for the standard condition, so it seemed reasonable to test whether results like these would be observed empirically. We therefore tested these model-based predictions in the three experiments reported next (Experiment 1, direct replication of Experiment 1, and Experiment 2). In the standard condition of these experiments, eyewitnesses reported their confidence in lineup rejections (as usual) without the researchers revealing the suspect. In the reveal condition of all experiments, if the participant rejected the lineup, the suspect was revealed before confidence was assessed.

## Method

We conducted three mock-crime studies to test the predictions illustrated in Figure 2: Experiment 1, involving a six-person lineup; a direct replication of Experiment 1; and Experiment 2, involving a nine-person lineup. In these studies, participants first watched a video of a simulated crime and were then presented with either a TP or TA simultaneous lineup. In the standard condition, confidence in a lineup rejection was not gathered in relation to a specific face, whereas in the reveal condition of all studies, the suspect was

revealed after the lineup was rejected and the participant was asked how sure they were that this person was *not* the perpetrator from the video. The methodological differences across studies are minor, but for clarity, we present separate method sections for the three studies.

## Experiment 1

### Participants

We recruited participants from Amazon’s Mechanical Turk (MTurk). We planned to test 1,000 participants because that sample size yielded approximately 80% power to detect a difference between high-confidence accuracy scores of .75 versus .90. Those values came from our simulation, which yielded high-confidence accuracy scores of  $\sim .76$  (*standard*) and  $\sim .89$  (*reveal*). Although the power analysis indicated that we needed to test approximately 1,000 subjects, in practice, the study ended up with a few more participants. Because we did not know in advance how many participants would need to be excluded due to not satisfying our inclusion criteria, we followed our standard protocol of keeping the study active on MTurk in 12-day blocks and checking to see if we had a sufficient number of usable participants after each block. If not, we ran another block until we achieved at least the minimum number of participants, at which point we analyzed the data.

Experiment 1 tested 1,117 participants, with 89 excluded because they incorrectly answered an attention check question at the end of the task or stated that they had seen the stimulus video before the study. The attention check consisted of two multiple-choice questions asking what crime took place in the video and where the crime took place. This resulted in a total number of 1,028 participants in the final analysis ( $M_{\text{age}} = 33.2$  years,  $SD = 10.2$  years; females = 543, males = 479, not reported = 6). The race breakdown of the final participants was as follows: White = 670, Asian = 141, Black = 69, Latino = 78, Native American = 32, other = 23, no response = 15. A total of 505 participants were randomly assigned to the standard condition (279 to a TP lineup and 226 to a TA lineup) and 523 were randomly assigned to the reveal condition (253 to a TP lineup and 270 to a TA lineup).

Participants were compensated 25 cents for the approximately 5 min required to complete the test. The study was approved by the University of California San Diego Institutional Review Board for



research involving human subjects (121186SX), and all participants provided informed consent before participating.

### Materials

The stimulus was a 30-s mock-crime video showing a male perpetrator painting graffiti on a classroom wall. In the video, the perpetrator's face is visible from the front and profile view for approximately 8 s, with all shots being close to the camera.

Our filler pool consisted of 60 filler photos taken from the Florida Department of Corrections, and they were judged to be description matched by an independent group of participants. More information about these ratings is included in the [Supplemental Material](#).

### Design and Procedure

After participants consented to participate and read the instructions, the experiment started with the presentation of a mock-crime video. After the video, participants played a visual matching-tile game for 45 s. When the game ended, participants entered the lineup phase of the experiment. The instructions indicated that the person from the video may or may not be in the lineup and that an ID did not have to be made because there would be a "not present" option available.

The lineup phase was a randomized 2 (control vs. experimental)  $\times$  2 (TP vs. TA) design. Participants were shown a six-person simultaneous lineup in which the photographs were arranged in two rows of three photographs. In the TP condition, one of the photographs in the lineup was the suspect from the video, whereas the remaining photographs were randomly drawn from a pool of fillers. In the TA condition, all of the photographs in the lineup were randomly drawn from the pool of fillers.

In the standard condition, participants could either choose one of the photos in the lineup as being the perpetrator from the video or select the "not present" button at the base of the lineup. After making their decision, participants were provided a confidence rating scale (0%–100%) and were prompted with the question "How certain are you?" The confidence rating was entered using a slider at the bottom of the screen. The method for gathering confidence was the same regardless of whether the participant chose the suspect photo or a filler photo.

The reveal condition was identical to the standard condition except in the case of lineup rejections. As before, if a participant made a positive ID of a suspect photo or a filler photo, confidence was gathered in the same manner described above. However, in this condition, if the participant rejected the TA or TP lineup, the suspect in the lineup was revealed to the participant before confidence was gathered. In other words, for TP lineups that were rejected, the perpetrator was identified, and for TA lineups that were rejected, a randomly selected filler serving as the designated innocent suspect was revealed to the witness. In each case, an enlarged pop-up photo of the suspect was superimposed on the lineup. The participant was then prompted to provide a confidence rating (0%–100%) indicating how sure they were that the highlighted individual was not the perpetrator from the video. Thus, participants were *not* given an opportunity to change their mind and make an ID. Instead, they were only asked to express confidence in their lineup rejection.

There were two attention check questions. The first was, "Where did the crime occur?" Possible answers were "in a classroom," "at a

park," "at the zoo," and "on a street." The second question was, "What was the person in the video doing?" Possible answers were "talking on a phone," "painting graffiti," "eating a sandwich," and "reading a book." Failure to correctly answer both questions resulted in exclusion from the final analysis.

### Preregistered Replication of Experiment 1

As noted later in the Results, Experiment 1 yielded mostly interpretable results but one apparent anomaly. Thus, to test the robustness of the findings of Experiment 1, we ran a preregistered replication (chronologically, this experiment was conducted after Experiments 1 and 2 were completed).

### Participants

We recruited participants from MTurk. We tested a total of 1,915 participants; of those, 535 were excluded using the same criteria used in Experiment 1 (407 because they had previously seen the stimulus video and 128 because they failed the attention check). This resulted in a total of 1,380 participants in the final analysis ( $M_{\text{age}} = 32.94$  years,  $SD = 10.35$  years; females = 647, males = 724, not reported = 9). The race breakdown of the final participants was as follows: White = 861, Asian = 205, Black = 95, Latino = 120, Native American = 28, other = 44, no response = 27. A total of 693 participants were randomly assigned to the standard condition (352 to a TP lineup and 341 to a TA lineup), and 687 were randomly assigned to the reveal condition (339 to a TP lineup and 348 to a TA lineup). By the time this replication study was conducted, the stimulus video had been used in several other studies, which is presumably why a larger number of participants reported having seen the video before.

Participants were compensated 50 cents for the approximately 5 min required to complete the test. The study was approved by the University of California San Diego Institutional Review Board for research involving human subjects (121186SX), and all participants provided informed consent before participating.

### Materials, Design, and Procedure

The rerun was a direct replication of Experiment 1.

### Experiment 2

Experiment 2 was nearly identical to Experiment 1 except that we increased the lineup size to nine. We did so because the confidence–accuracy relation for lineup rejections in Experiment 1 did not turn out to be as flat as we expected. Previous work reporting a flat function used a larger lineup size of eight (e.g., [Brewer & Wells, 2006](#)), leading us to speculate that the larger the lineup size, the flatter the confidence–accuracy function for lineup rejections would be.

### Participants

We recruited participants from MTurk. Experiment 2 tested 1,094 participants; of those, 156 were excluded using the same criteria used in Experiment 1. This resulted in a total of 938 participants in the final analysis ( $M_{\text{age}} = 32.2$  years,  $SD = 10.2$  years; female = 462, male = 468, not reported = 8). The race breakdown of the final participants was as follows: White = 624, Asian = 100, Black = 63,

Latino = 86, Native American = 12, other = 29, no response = 24. Only after analyzing the data did we realize that we had not quite reached the goal of 1,000 participants, but we decided to report the results on the assumption that they likely would not differ appreciably after 62 more participants were tested. Of the 938 usable participants, 502 were randomly assigned to the standard condition (259 to a TP lineup and 243 to a TA lineup) and 436 were randomly assigned to the reveal condition (225 to a TP lineup and 211 to a TA lineup).

Participants were compensated 25 cents for the approximately 5 min required to complete the test. The study was approved by the University of California San Diego Institutional Review Board for research involving human subjects (121186SX), and all participants provided informed consent before participating.

### Materials

The materials used in Experiment 2 were identical to those in Experiment 1.

### Design and Procedure

Experiment 2's design and procedure were identical to those of Study 1 with only these minor differences. Experiment 2 involved a nine-person simultaneous lineup in which the photographs were arranged in three rows of three photographs. TP lineups consisted of one suspect photo and eight randomly drawn filler photos. TA lineups consisted of nine filler photos. We wanted to test a nine-person lineup because although we obtained asymmetrical confidence–accuracy relationships for positive IDs versus lineup rejections in the standard condition of Experiment 1, we did not obtain the flat relationship for lineup rejections that has been previously observed (Brewer & Wells, 2006). Given that Brewer and Wells (2006) used eight-person lineups, we thought a larger lineup might weaken the confidence–accuracy relationship for lineup rejections because the confidence decision would be distributed across a larger set of faces (Lindsay et al., 2013). Experiment 2 used nine-person lineups (slightly larger than those used by Brewer & Wells, 2006) to test this idea.

In the reveal procedure, guilty and designated innocent suspects were revealed with a box around their photo instead of an enlarged pop-up photo. This was not an intentional change in the operationalization of the reveal procedure. The difference is due to ease of programming for different programmers working on the project.

### Results

Table 1 shows the overall suspect ID rates, filler ID rates, and lineup rejection rates from TP lineups and TA lineups for both the standard and reveal conditions in each experiment. The data appear unremarkable except that in Experiment 1, the TA lineup rejection rates for the standard and reveal conditions differed considerably (.67 and .77, respectively). They should have been equivalent because the procedures were identical up to the point of a lineup rejection. Because there was no obvious reason why that difference occurred, we replicated the experiment to test for the robustness of the main results. In the replication, the key findings were similar, but the anomaly was no longer apparent.

### Calibration Plots

We next analyzed the data broken down by confidence. As in past research (e.g., Mickes, 2015; Wilson et al., 2018), we aggregated confidence ratings into three bins: 0–69, 70–89, and 90–100. Doing so created bins with approximately equal numbers of observations. The top row of Figure 3 shows the calibration plots from the standard condition for Experiment 1, Experiment 1's replication, and Experiment 2. These calibration accuracy scores were computed as they typically are and as they were for the simulated data presented earlier.

In the standard condition (top row of Figure 3), the slope for lineup rejections visually appears to be somewhat shallower than the slope for positive IDs. However, the asymmetry between the slopes is not as large as that observed in some other studies. For example, Brewer and Wells (2006) reported flat confidence–accuracy relationships for lineup rejections in some—but not all—of the conditions they ran (see Figure 2 of that article). In our experiments, the lineup rejections were reasonably diagnostic of accuracy. This remained true even when the lineup size increased to nine in Experiment 2 (top right of Figure 3, open symbols). In the reveal condition, the slopes of the confidence–accuracy plots for positive IDs versus lineup rejections are visually similar (bottom row of Figure 3).

### Lineup Rejection Accuracy

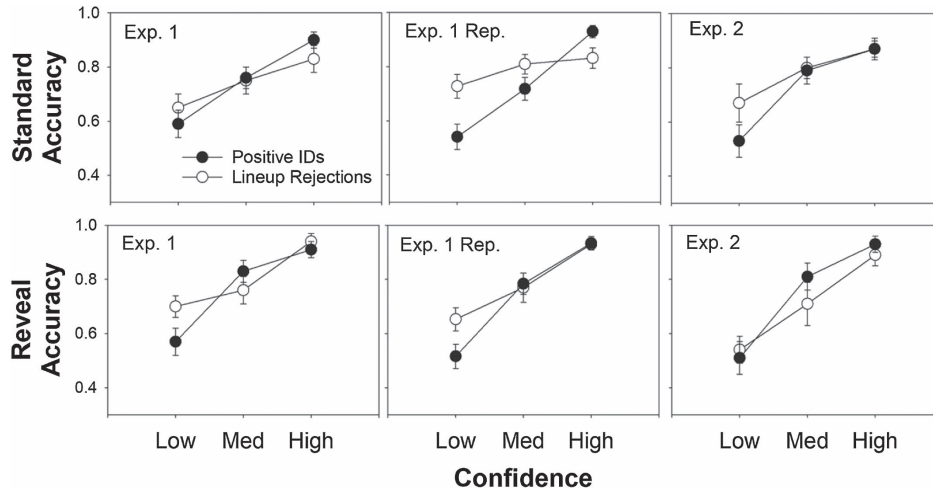
Figure 4 directly compares the confidence–accuracy functions for lineup rejections from the standard and reveal conditions (plotted separately in Figure 3). In Experiment 1, the accuracy of high-confidence lineup rejections in the reveal condition (.94) was considerably higher than the accuracy of high-confidence lineup

**Table 1**

*Suspect ID, Filler ID, and No ID Rates for Target-Present and Target-Absent Lineups in the Standard and Reveal Conditions*

Experiment	Condition	Target present			Target absent	
		Suspect	Filler	No ID	Filler	No ID
Experiment 1	Standard	.77	.05	.18	.33	.67
	Reveal	.78	.05	.17	.23	.77
Experiment 1 replication	Standard	.75	.07	.19	.27	.73
	Reveal	.73	.07	.19	.26	.74
Experiment 2	Standard	.71	.08	.20	.31	.69
	Reveal	.68	.09	.23	.29	.71

**Figure 3**  
*Confidence–Accuracy Calibration Plots for Low-, Medium-, and High-Confidence Positive IDs and Lineup Rejections*



*Note.* Confidence–accuracy calibration plots for the standard (top row) and reveal (bottom row) conditions. The procedure for positive IDs (results shown using filled symbols) was the same in both conditions of all three experiments, but the procedure for lineup rejections (results shown using open symbols) differed between the standard and reveal conditions.

rejections in the standard condition (.83), and this difference was significant,  $\chi^2(1) = 6.35, p = .012, \phi = .18$ . Experiment 1’s replication yielded similar results; the accuracy of high-confidence lineup rejections was significantly higher in the reveal condition (.93) than in the standard condition (.83),  $\chi^2(1) = 5.60, p = .018, \phi = .15$ . In Experiment 2, accuracy was again higher in the reveal condition (.89 vs. .87), but the difference was not significant,  $\chi^2(1) = 0.16, p = .690, \phi = .033$ .

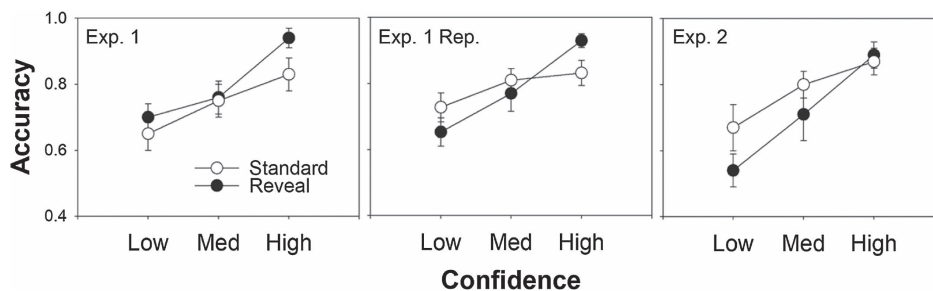
**Frequency of High-Confidence Rejections**

Next, we examined the proportion of TA and TP lineups associated with a high-confidence lineup rejection. As shown in Figure 5, the proportion of innocent suspects in TA lineups associated with high-confidence lineup rejections was substantially higher in the

reveal condition than in the standard condition. The difference was significant in Experiment 1,  $\chi^2(1) = 9.97, p < .002, \phi = .14$ ; in Experiment 1’s replication,  $\chi^2(1) = 17.67, p < .001, \phi = .16$ ; and in Experiment 2,  $\chi^2(1) = 39.12, p < .001, \phi = .29$  (Figure 5).

By contrast, the proportion of guilty suspects associated with high-confidence lineup rejections did not differ across the standard versus reveal conditions for any of the experiments—Experiment 1:  $\chi^2(1) = 1.31, p = .25, \phi = .05$ ; Experiment 1 replication:  $\chi^2(1) = 1.21, p = .27, \phi = .04$ ; Experiment 2:  $\chi^2(1) = 2.35, p < .13, \phi = .07$ . We also calculated the Bayes factors for these nonsignificant changes in the number of guilty suspects associated with high-confidence rejections. The odds in favor of the null of no difference were 14.24 for Experiment 1, 17.06 for Experiment 1’s replication, and 8.10 for Experiment 2. Taken together, the findings summarized in Figures 4 and 5 suggest that revealing the suspect to witnesses *after* they reject

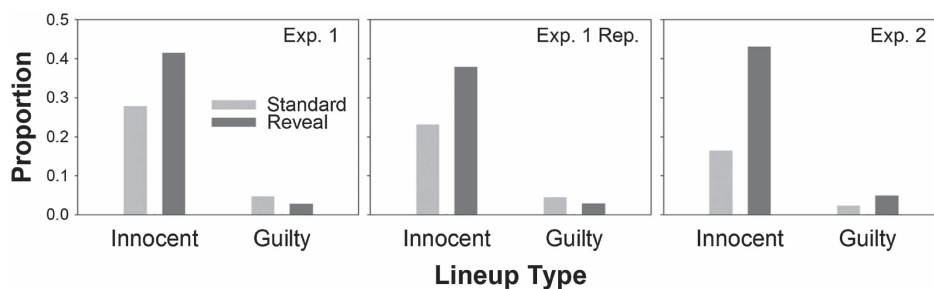
**Figure 4**  
*Confidence–Accuracy Plots for Low-, Medium-, and High-Confidence Lineup Rejections*



*Note.* Confidence–accuracy plots for lineup rejections from the standard condition (open symbols) and for suspect rejections from the reveal condition (filled symbols).



**Figure 5**  
Suspects Receiving High-Confidence Rejections



*Note.* Proportion of innocent suspects (in target-absent lineups) and guilty suspects (in target-present lineups) receiving high-confidence rejections.

the lineup but *before* assessing their confidence could enhance evidence of innocence and potentially clear many more innocent suspects.

### General Discussion

We tested predictions from a signal detection model about how to enhance evidence of innocence when a simultaneous lineup is rejected. The model assumes that, whereas confidence in a positive ID is based on the memory signal associated with the identified face, confidence in a lineup rejection is based on the average memory signal of all faces in the lineup. With its parameters set to yield asymmetrical confidence–accuracy relationships for positive IDs versus lineup rejections, the model predicted that two types of information could be enhanced in the case of lineup rejections by implementing one slight change: revealing the suspect to the witness after they make a lineup decision but before collecting confidence ratings. Specifically, the model predicted that, using the reveal procedure, (a) the accuracy of high-confidence rejections would increase and (b) the number of innocent suspects who received a high-confidence rejection would also increase (Figure 2).

The results of the three experiments reported here were generally consistent with these predictions. Across all three experiments, the reveal procedure increased the number of innocent suspects receiving a high-confidence rejection. In Experiment 1 and its replication, there was also a significant increase in the accuracy of high-confidence rejections. In Experiment 2, the accuracy of high-confidence rejections in the reveal condition also increased but not significantly so. There are several possible reasons for that lack of significance in Experiment 2. First, it could simply be a Type II error (i.e., even real differences are not always statistically significant). For example, assuming a true increase in accuracy for the reveal condition relative to the standard condition, the probability of obtaining a significant result in all three experiments, each with 80% power, is only  $.8^3 \approx .5$  (i.e., one would not expect to see a significant result every time the experiment is run even when power is adequate). From this perspective, the overall pattern of results across the three experiments suggests that the accuracy of high-confidence rejections is in fact higher for the reveal condition. Alternatively, the null hypothesis of no difference might be true for Experiment 2, which might mean that the reveal procedure’s effectiveness differs for six-person and nine-person lineups. This interpretation seems unlikely to us because we are aware of no theoretical reason to believe it might be true. Still, we cannot rule it out.

### Theoretical Implications

Confidence in a positive ID is presumably based on the strength of the memory signal generated by the face in the lineup that yields the MAX signal (i.e., the decision variable is the MAX signal). We adopted the additional assumption that the decision variable differs for lineup rejections because of the flatter (sometimes completely flat) confidence–accuracy relationship that is often observed for nonchoosers. A seemingly reasonable alternative decision variable for determining confidence in a lineup rejection is the average memory signal associated with the faces in the lineup (e.g., Lindsay et al., 2013).

The findings reported here lend some credence to this idea. At least using the specific parameters that generated the simulated results shown in Figure 2, a signal detection model incorporating that idea can yield asymmetrical confidence–accuracy relationships for positive IDs versus lineup rejections (which was not previously known), and it can generate predictions for the reveal condition that correspond to what we found here. Thus, the notion that lineup rejections are based on an average memory signal seems viable. Whether the model makes these predictions for all possible parameter settings is not known, but it does so for the parameters we investigated.

Of course, there are other possibilities that we did not investigate and that might also be viable. For example, the decision variable for lineup rejections might be the sum of the memory signals generated by the faces in the lineup, or it might instead be based on a subset of the faces in the lineup (e.g., the average or the sum of the two faces in the lineup that generate the strongest memory signal). Although we cannot rule out alternative models such as those, our results do indicate that the model we investigated, which assumes that the decision variable for lineup rejections is an average signal, is viable. Not only can it account for the asymmetrical confidence–accuracy relationships that have been observed for positive IDs versus lineup rejections, it also correctly predicted that the accuracy of high-confidence lineup rejections (and the number of innocent suspects receiving a high-confidence rejection) would increase in the reveal condition compared to the standard condition.

### Applied Implications

If the results reported here are confirmed by follow-up research, the reveal procedure has the potential to quickly remove suspicion of guilt from innocent suspects on the first lineup test. Thus, its use in actual police lineups could help to reduce the amount of contact the

innocent have with the legal system. This would be beneficial not only because it would help to prevent wrongful convictions but also because “contact with the justice system—through spending a night in jail/prison, being issued an ASBO [antisocial behavior order], or having an official crime record—promotes misbehavior” (Motz et al., 2020, p. 323). Increased contact with the criminal justice system also raises the chances of being labeled as a criminal in the future (Motz et al., 2020). A guilt-clearing high-confidence lineup rejection would allow law enforcement to quickly shift their suspicion to a new (potentially guilty) suspect.

Implementing the reveal procedure in practice would be easy to do in many police departments but would be more challenging in others. For example, in approximately 30% of police departments in the United States, the lineup administrator does not know the identity of the suspect (Police Executive Research Forum, 2013). How, then, could the reveal procedure be implemented? There are two feasible options for those departments. One option worth considering is a hybrid approach in which everyone in the lineup receives a confidence rating (cf. Brewer et al., 2020), but only after the lineup is rejected. Under this hybrid method, the suspect photo would still receive a confidence rating, as in a reveal procedure, but both administrators and witnesses could remain blind to the suspect’s identity. For positive IDs, no hybrid approach would be needed, and current procedures would remain in effect. More empirical support is needed before implementation, however, given that we did not investigate a hybrid approach compared to other lineups.

Another option would be to use a computerized lineup. According to the Police Executive Research Forum (2013), approximately 12% of police departments administer lineups using a computer. A computer-based approach allows for the implementation of a reveal procedure while still maintaining a double-blind, neutral lineup. We want to further emphasize that we are not advocating for the abandonment of double-blind lineup administration. Double-blind lineups are considered “best practice” for administering lineups and photospreads (Wells et al., 2020). We advocate for more research on how to best refine the reveal procedure so that innocent suspects are not unnecessarily imperiled. Regardless of how the reveal procedure is implemented (and as long as it is a properly administered, fair lineup), our main point is that there is considerable value in obtaining a confidence rating for the suspect in the lineup after the lineup is rejected.

## Limitations

In light of the findings from these experiments, should the police immediately change how they conduct lineups? We do not think so. First, it would be important for independent labs to directly replicate these findings in large-*N* studies to ensure reliability (Wilson et al., 2020). In addition, our lineup experiments all used the same stimulus materials, so replications with varied stimuli are important before recommending implementation (e.g., Wells & Windschitl, 1999). In fact, “radical randomization” of variables that theoretically should not affect the conclusions (e.g., age, race, or gender of the perpetrator and lineup members) is an important step in ensuring generalizability (Baribault et al., 2018). There is no theoretical reason to believe that the findings reported here are specific to the stimulus materials we used, but the possibility cannot be ruled out. Therefore, testing generalizability is essential before implementing the reveal procedure in practice. Second, it would be important to also test the revised

lineup procedure in a police setting to ensure generalizability to the real world.

Finally, it is important that lab-based research conduct these tests with diversity of participants in mind. Although our sample was more diverse than what is often found in the psychological sciences (Gosling et al., 2010), we still had a majority White population. A “lack of racial diversity in psychology stands to leave the field unprepared for an increasingly diverse society” and may result in “theories, methods, and findings that do not reflect a diversity of perspectives” (Roberts et al., 2020, p. 1304). It is important to consider that lab-based studies often do not consider the social contexts of the real world (Kovera & Evelo, 2021). However, if our findings can be independently replicated—and if they generalize to other stimulus materials and to the police setting—policymakers should seriously consider implementing procedural changes that, compared to current procedures, would clear substantially more innocent suspects early in a police investigation.

## Future Directions

There are several important issues to sort out in future research. For example, what conditions determine whether the confidence–accuracy relationship for lineup rejections is flat (as has sometimes been observed in previous research) versus being somewhat diagnostic of accuracy? The reveal procedure appears to reliably increase the proportion of innocent suspects receiving a high-confidence lineup rejection. If lineup rejections are at least somewhat diagnostic of accuracy (as they were in all three experiments reported here), then more innocent suspects receiving a more accurate rejection would be an undeniably positive outcome. However, it would not be a positive outcome if (a) the confidence–accuracy relationship for lineup rejections is completely flat (as has sometimes been observed) and (b) the reveal procedure does not increase the accuracy of a lineup rejection. In our study, the reveal procedure did significantly increase the accuracy of high-confidence lineup rejections in two out of three experiments, but it will be important to determine if that result is observed when the confidence–accuracy relationship for lineup rejections in the standard condition is completely flat.

## Conclusion

The problem of lineup rejections not being especially diagnostic of innocence has been apparent for a long time. However in recent years, it has become clear that under certain conditions, positive IDs made with high confidence can be highly diagnostic of guilt (Wixted & Wells, 2017). Although much work remains to be done, the reveal procedure appears to offer a promising new approach for making high-confidence rejections also highly diagnostic of innocence.

## References

- Baribault, B., Donkin, C., Little, D. R., Trueblood, J. S., Oravecz, Z., van Ravenzwaaij, D., White, C. N., De Boeck, P., & Vandekerckhove, J. (2018). Metastudies for robust tests of theory. *Proceedings of the National Academy of Sciences of the United States of America*, 115(11), 2607–2612. <https://doi.org/10.1073/pnas.1708285114>
- Brewer, N., Weber, N., & Guerin, N. (2020). Police lineups of the future?. *American Psychologist*, 75(1), 76–91. <https://doi.org/10.1037/amp0000465>

- 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*(1), 11–30. <https://doi.org/10.1037/1076-898X.12.1.11>
- Cutler, B. L., Penrod, S. D., & Martens, T. K. (1987). The reliability of eyewitness identifications: The role of system and estimator variables. *Law and Human Behavior*, *11*(3), 233–258. <https://doi.org/10.1007/BF01044644>
- Gosling, S. D., Sandy, C. J., John, O. P., & Potter, J. (2010). Wired but not WEIRD: The promise of the internet in reaching more diverse samples. *Behavioral and Brain Sciences*, *33*(2–3), 94–95. <https://doi.org/10.1017/S0140525X10000300>
- Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. Wiley.
- Kovera, M. B., & Evelo, A. (2021). Eyewitness identification in its social context. *Journal of Applied Research in Memory and Cognition*, *10*(3), 313–327. <https://doi.org/10.1016/j.jarmac.2021.04.003>
- Lee, J., & Penrod, S. D. (2019). New signal detection theory-based framework for eyewitness performance in lineups. *Law and Human Behavior*, *43*(5), 436–454. <https://doi.org/10.1037/lhb0000343>
- Lindsay, R. C. L., Kalmel, N., Leung, J., Bertrand, M. I., Sauer, J. D., & Sauerland, M. (2013). Confidence and accuracy of lineup selections and rejections: Postdicting rejection accuracy with confidence. *Journal of Applied Research in Memory and Cognition*, *2*(3), 179–184. <https://doi.org/10.1016/j.jarmac.2013.06.002>
- Lindsay, R. C. L., & Wells, G. L. (1985). Improving eyewitness identification from lineups: Simultaneous versus sequential lineup presentations. *Journal of Applied Psychology*, *70*(3), 556–564. <https://doi.org/10.1037/0021-9010.70.3.556>
- Mickes, L. (2015). Receiver operating characteristic analysis and confidence-accuracy characteristic analysis in investigations of system variables and estimator variables that affect eyewitness memory. *Journal of Applied Research in Memory and Cognition*, *4*(2), 93–102. <https://doi.org/10.1016/j.jarmac.2015.01.003>
- Mickes, L., Flowe, H. D., & Wixted, J. T. (2012). Receiver operating characteristic analysis of eyewitness memory: Comparing the diagnostic accuracy of simultaneous versus sequential lineups. *Journal of Experimental Psychology: Applied*, *18*(4), 361–376. <https://doi.org/10.1037/a0030609>
- Motz, R. T., Barnes, J. C., Caspi, A., Arseneault, L., Cullen, F. T., Houts, R., Wertz, J., & Moffitt, T. E. (2020). Does contact with the justice system deter or promote future delinquency? Results from a longitudinal study of British adolescent twins. *Criminology*, *58*(2), 307–335. <https://doi.org/10.1111/1745-9125.12236>
- Police Executive Research Forum. (2013). *A national survey of eyewitness identification procedures in law enforcement agencies*. <http://policeforum.org/library/eyewitness-identification/NIJEyewitnessReport.pdf>
- Roberts, S. O., Bareket-Shavit, C., Dollins, F. A., Goldie, P. D., & Mortenson, E. (2020). Racial inequality in psychological research: Trends of the past and recommendations for the future. *Perspectives on Psychological Science*, *15*(6), 1295–1309. <https://doi.org/10.1177/1745691620927709>
- Sauerland, M., Sagana, A., & Sporer, S. L. (2012). Assessing nonchoosers' eyewitness identification accuracy from photographic showups by using confidence and response times. *Law and Human Behavior*, *36*(5), 394–403. <https://doi.org/10.1037/h0093926>
- Smith, A. M., Lampinen, J. M., Wells, G. L., Smalarz, L., & Mackovichova, S. (2019). 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*, *8*(1), 50–59. <https://doi.org/10.1016/j.jarmac.2018.09.003>
- Smith, A. M., Yang, Y., & Wells, G. L. (2020). Distinguishing between investigator discriminability and eyewitness discriminability: A method for creating full receiver operating characteristic curves of lineup identification performance. *Perspectives on Psychological Science*, *15*(3), 589–607. <https://doi.org/10.1177/1745691620902426>
- Wells, G. L., Kovera, M. B., Douglass, A. B., Brewer, N., Meissner, C. A., & Wixted, J. T. (2020). Policy and procedure recommendations for the collection and preservation of eyewitness identification evidence. *Law and Human Behavior*, *44*(1), 3–36. <https://doi.org/10.1037/lhb0000359>
- Wells, G. L., & Olson, E. A. (2002). Eyewitness identification: Information gain from incriminating and exonerating behaviors. *Journal of Experimental Psychology: Applied*, *8*(3), 155–167. <https://doi.org/10.1037/1076-898X.8.3.155>
- Wells, G. L., & Windschitl, P. D. (1999). Stimulus sampling and social psychological experimentation. *Personality and Social Psychology Bulletin*, *25*(9), 1115–1125. <https://doi.org/10.1177/01461672992512005>
- Wells, G. L., Yang, Y., & Smalarz, L. (2015). Eyewitness identification: Bayesian information gain, base-rate effect equivalency curves, and reasonable suspicion. *Law and Human Behavior*, *39*(2), 99–122. <https://doi.org/10.1037/lhb0000125>
- Wilson, B. M., Harris, C. R., & Wixted, J. T. (2020). Science is not a signal detection problem. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(11), 5559–5567. <https://doi.org/10.1073/pnas.1914237117>
- Wilson, B. M., Seale-Carlisle, T. M., & Mickes, L. (2018). The effects of verbal descriptions on performance in lineups and showups. *Journal of Experimental Psychology: General*, *147*(1), 113–124. <https://doi.org/10.1037/xge0000354>
- Wixted, J. T., Vul, E., Mickes, L., & Wilson, B. M. (2018). Models of lineup memory. *Cognitive Psychology*, *105*, 81–114. <https://doi.org/10.1016/j.cogpsych.2018.06.001>
- Wixted, J. T., & Wells, G. L. (2017). The relationship between eyewitness confidence and identification accuracy: A new synthesis. *Psychological Science in the Public Interest*, *18*(1), 10–65. <https://doi.org/10.1177/1529100616686966>

Received September 8, 2020

Revision received October 25, 2021

Accepted November 6, 2021 ■