Response Bias Modulates the Confidence-Accuracy Relationship for both Positive Identifications and Lineup Rejections in a Simultaneous Lineup Task

Authors: Anne S. Yilmaz, Xiaoqing Wang, John T. Wixted

University of California, San Diego

The authors have no conflicts of interest to declare.

Author Note

Correspondence concerning this article should be addressed to John Wixted, Department of Psychology, University of California, San Diego, La Jolla, CA, 92093. orcid ID: 0000-0001-6282-5479. Email: jwixted@ucsd.edu. Supported by a grant from the UCSD Yankelovich Center and in part by the Center of Academic Research and Training in Anthropogeny (CARTA) Fellowship.

Abstract

In recent years, the use of calibration analysis and confidence-accuracy characteristic analysis has revealed the confidence-accuracy relationship for positive identification (ID) made from a lineup is often strong. At the same time, the confidence-accuracy relationship for lineup rejections is typically much weaker. Why the relationship is often weak for lineup rejections remains unclear. Here, we report two experiments testing a prediction that follows from signal detection theory. Specifically, this theory predicts that one determinant of the strength of the confidence-accuracy relationship for both positive IDs and lineup rejections is response bias. Theoretically, inducing a more conservative response bias should weaken the confidence-accuracy relationship for positive IDs while strengthening it for lineup rejections. The two experiments reported here support this prediction.

Key words: Confidence-Accuracy Relationship; Lineup Rejections; Signal Detection Theory

Response Bias Modulates the Confidence-Accuracy Relationship for both Positive Identifications and Lineup Rejections in a Simultaneous Lineup Task

Eyewitness memory is often tested using a lineup consisting of one suspect (who is either innocent or guilty) and five or more physically similar fillers. A witness can either make a positive identification (picking either the suspect or a filler) or reject the lineup altogether. A key question that the field has addressed for over 40 years concerns the confidence in a positive identification from a lineup and the accuracy of that identification. Interest in this question can be traced to the many high-confidence identifications made at criminal trials that were shown to be incorrect when the convicted defendant was ultimately exonerated by DNA evidence. However, our focus here is on the confidence-accuracy relationship on the first test of a witness's memory (e.g., using a lineup), not the last test conducted at trial, often a year or two later.

The field once concluded that, even on an initial and properly administered lineup, confidence was, at best, only weakly related to accuracy. However, over time, it has become increasingly clear that the opposite is true (Juslin et al., 1996; Brewer & Wells, 2006; Wixted et al., 2015; Wixted & Wells, 2017). In fact, for positive identifications of the suspect from a pristine lineup (i.e., for the subset of eyewitnesses who pick the suspect), confidence is strongly predictive of accuracy in the sense that high-confidence identifications are highly accurate and low-confidence identifications are highly inaccurate (often close to chance). This is true even of actual eyewitnesses tested during a police investigation (Quigley-McBride & Wells, 2023; Wixted et al., 2016).

However, the strength of the confidence-accuracy relationship appears to be much less impressive when it comes to lineup rejections. Indeed, in contrast to the strong relationship for positive IDs, the relationship between confidence and accuracy for lineup rejections is often (but

not always) found to be negligible (Brewer & Wells, 2006; Arndorfer & Charman, 2022). Thus, a high-confidence lineup rejection is not necessarily indicative of high accuracy like it is in the case of a positive identification.

Although the field has already reached a de facto consensus about the nature of the confidence-accuracy relationship in the case of lineup rejections, no formal review of the past literature has been conducted in the manner previously done for positive IDs by Wixted & Wells (2017). We therefore did so here by reviewing the confidence-accuracy relationship for lineup rejections reported in 12 experiments (Brewer et al., 2002; Brewer & Wells, 2006; Carlson et al., 2017; Dobolyi & Dodson, 2013; Dodson & Dobolyi, 2016; Horry et al., 2012; Keast et al., 2007; Palmer et al., 2013; Sauer et al., 2008; 2010; Sauerland & Sporer, 2009; Weber & Brewer, 2004). Except for a few studies that did not report confidence for lineup rejections, these are the same experiments reviewed by Wixted and Wells (2017) to assess the confidence-accuracy relationship for positive IDs made using a 100-point confidence scale. The data sets span a variety of study designs, such as: simultaneous and sequential lineups, same-race and cross-race identifications, methodologies (e.g., disconfirmation and reflection, immediate presentation and delayed presentation, etc.), as well as different sample populations (e.g., adults and children). Figure 1 shows the average confidence-accuracy relationship for positive suspect IDs reported by Wixted and Wells (2017) and for lineup rejections. Clearly, the relationship is weaker for lineup rejections.

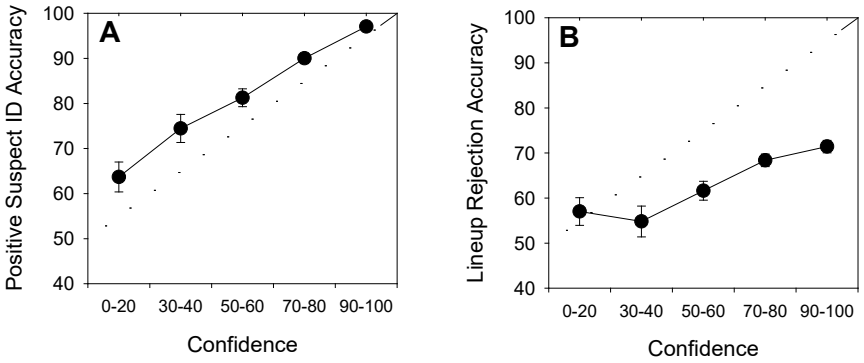


Figure 1. (A) Confidence-accuracy characteristic for positive IDs reported by Wixted and Wells (2017). (B) Confidence-accuracy characteristic for lineup rejections from the same studies that reported confidence for lineup rejections.

The question of interest here is why there is an asymmetry between the confidence-accuracy relationship for positive vs. lineup rejections. Picking up on an idea suggested by Brewer and Wells (2006) and Lindsay et al. (2013), Yilmaz et al. (2022) hypothesized that confidence in lineup rejections might be determined by the average memory signal because no singular face is identified when a lineup is rejected. This is in contrast to positive IDs, where confidence is presumably based on the one that generates the strongest memory-match signal (the MAX face). However, using a model-fitting approach across six different data sets, Yilmaz and Wixted (submitted) found that confidence in a lineup rejection also appears to be based on the MAX face (i.e., the less familiar the MAX is, the more confidence the witness is in rejecting the lineup).

Smith et al. (2023) hypothesized that a focus on suspect IDs for positive IDs may explain the asymmetry in the confidence-accuracy relationship for positive vs. lineup rejections. Unlike for positive IDs, for lineup rejections, an outcome is counted as correct or incorrect whether the MAX signal is generated by the suspect or a filler because, when a lineup is rejected, it is not known which face generated the MAX signal. Moreover, the distribution of memory-match

signals for innocent and guilty suspects overlap to a lesser degree (i.e., discriminability is higher) compared to the distribution of MAX memory-match signals (regardless of whether the MAX face is the suspect or a filler). However, for positive IDs, the confidence-accuracy relationship is not appreciably affected over a fairly wide range of discriminability, so it is not clear that this factor would explain the asymmetry.

Here, we investigate the possibility that the asymmetry might be explained, at least in part, based on the relatively high overall choosing rates (liberal response bias) observed in many lineup studies. One can conceptualize response bias in a police lineup as a witness' willingness to select a person as being the perpetrator. A liberal witness is more likely to select a face as being the guilty person (suspect or filler), while a conservative witness is more likely to reject the lineup. Within a signal detection framework, if participants have a liberal response bias, the decision criterion shifts to the left (Figure 2). This leftward shift means that lower degrees of memory strength are likely to surpass the decision criterion, thereby causing the witness to report a memory match. This increases the number of correct IDs (e.g., "hits") and well as false IDs (e.g., "false alarms," including false IDs of the innocent suspect and innocent fillers). Conversely, a conservative response bias causes the decision criterion to shift to the right, making it less likely that a witness reports a memory match. With increasing levels of conservatism, increasingly higher levels of memory strength are required for a witness to report a person as being the perpetrator (i.e., lowering both the hit rate and false alarm rate).

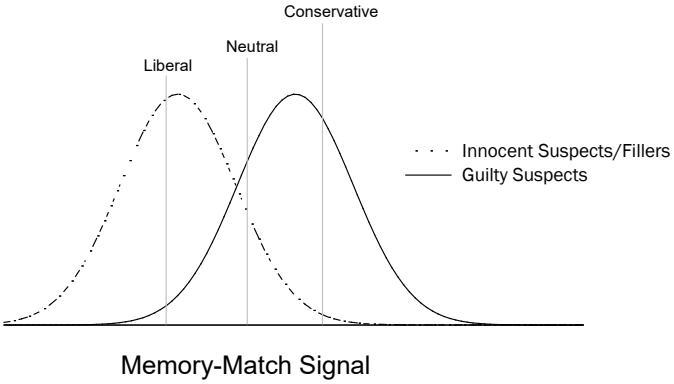


Figure 2. A standard signal detection model illustrating different place of the overall decision criterion for making positive IDs (liberal, neutral, and conservative).

Response bias may help account for the difference in the shape of the confidence-accuracy characteristic (CAC) that is often observed for positive IDs and lineup rejections. Specifically, liberal responding allows for a wider range of memory signal strengths to be the basis of confidence for positive IDs since more of each distribution exists above the decision criterion. The wider range allows for a steeper CAC for positive IDs. Reciprocally, if liberal responding increases the range of possible memory signal strengths associated with making an identification through the shifting of the decision criterion to the left, that shift would also *decrease* the range of possible memory signal strengths associated with a lineup rejection. Range restriction could explain why there is often little relationship between confidence and accuracy for lineup rejections — it could cause the slope of the CAC to flatten as there is less of each distribution falling to the left of the decision criterion. This logic would extend to conservative response biases as well. Shifting the decision criterion to the right should decrease the range of memory strengths associated with a positive ID, and expand the range of memory signals associated with a lineup rejection.

In eyewitness research, the focus is often on finding ways to induce more conservative responding for witnesses as it reduces the likelihood of a misidentification (Clark, 2005). Common examples of this focus are exemplified by the recommendations that witnesses should be informed that the guilty person may not be in the lineup, and that they have the option of rejecting the lineup if they don't believe the perpetrator is present (Technical Working Group for Eyewitness Evidence, 1999; Wells et al., 2020). Even with such instructions, overall choosing rates may be sufficiently high (response bias sufficiently liberal) that it may allow for a wide range accuracy associated with low to high confidence. Here, we hypothesize that although a liberal response bias will correspond to a relatively flat confidence-accuracy relationship for lineup rejections (as is typically observed), a conservative response bias for positive IDs will correspond to a steeper confidence-accuracy relationship for lineup rejections.

Experiment 1

In Experiment 1, we manipulated response bias using lineup instructions (liberal vs. conservative) to assess its effect on the confidence-accuracy relationship for positive IDs and lineup rejections.

Method

Participants

We recruited participants from Amazon's MTurk (n = 2,250). All participants passed attention check questions and reported that they had not seen the stimulus video before. Participants were compensated 25 or 50 cents for their time. The participants included 42.8% Male (1006), 52% Female (1222), 0.25% Other (6), 0.08% Decline to Answer (2), and 4.85% no response (114). The ethnicity distribution of the participants was: 82.5% Caucasian (1,939), 3.14% African-American (74), 8.42% Asian (198), 3.19% Latino (75), 0.5% Native-American

(12), 0.26% Middle-Eastern (6), 0.12% Pacific-Islander (3), 1% Other (23), 0.6% Decline to Answer (13), and 0.3% No Response (7).

Design and Materials

We used a randomized 2 (liberal vs. conservative instructions) x 2 (target present vs. target absent) design.

Procedure

The experiment started with a 24-second mock crime video. In the video, a man walks down a hallway in an office building and notices a laptop sitting unattended within a nearby office. The man looks arounds, enters the office, steals the laptop and walks away briskly. After the stimulus video, participants did a 45-second visual distractor task and then moved to the lineup phase.

For the instructions of the lineup phase in Experiment 1, participants were first told: “Imagine you are participating in a real police investigation, and the video you watched showed a real perpetrator committing a real crime. On the next page, you will be presented with some photos (also known as a "lineup"). The lineup may or may not contain the perpetrator of the crime you witnessed. If the perpetrator is present, click on his face. If he is NOT present, click the "Not Present" button. Regardless of your choice, you will then be asked for your confidence level ranging from 1-100. On the next screen, you will receive very important instructions along with the lineup. Please follow these instructions carefully.”

After clicking the “Next” button, participants received one of two lineup conditions: one with conservative instructions and one with liberal instructions. The conservative instructions read as follows: “IMPORTANT: These lineups almost never contain the photo of the perpetrator from the video. For this reason, it would be better to choose ‘Not Present’ than to select a face

and be wrong." The liberal instructions read as: "IMPORTANT: These lineups nearly always contain the photo of the perpetrator from the video. For this reason, it would be better to select a face and be wrong than to click 'Not Present.'"

The composition of the lineup itself (i.e., the photo array, not the lineup instructions) were the same regardless of condition. The lineup was a standard simultaneous lineup with two rows of three photographs. In the target present condition, one photo in the lineup was of the guilty suspect (i.e., the man from the video) while the other five photographs were fillers. Fillers are known-to-be-innocent faces included to help construct the lineup. The target absent condition did not contain a photo of the perpetrator. Instead, there was a sixth filler photo. Filler photos in the lineup were randomly selected from a pool of 60 possible fillers, all description-matched to the guilty suspect.

Participants could select a photograph as being the man from the video or they could reject the lineup by indicating that the man from the video was not present. After participants selected a face or rejected the lineup, they give their confidence (1%-100%; 1% = completely unsure; 100% = completely sure).

Both Experiment 1 and Experiment 2 were approved by the UCSD IRB (protocol # 121186), and the data we analyze here are available at https://osf.io/w8hnd/?view_only=bc0463105dac4b76819d8d63399a026c.

Results

The overall choosing rate from TP lineups in the liberal condition (suspect IDs plus filler IDs divided the number of TP lineups) was .88, whereas the corresponding value for the conservative condition was .74, a difference that was significant, $\chi^2 = 33.94, p < .001$. The overall choosing rate from TA lineups in the liberal condition (filler IDs divided the number of

TA lineups) was .42, whereas the corresponding value for the conservative condition was .26, a difference that was also significant, $\chi^2 = 33.65, p < .001$. In other words, choosing rates were significantly lower in the conservative condition for both TA and TP lineups, indicating that response bias was successfully manipulated.

Bins for low, medium, and high confidence were constructed such that each bin's frequency is roughly equated (i.e., 100-90 = high confidence; 89-70 = medium confidence; 69-1 = low confidence). This binning is typical, and the results discussed next are not appreciably affected by the choice of confidence bins. The frequency counts are shown in Table 1.

Table 1. Frequency counts by confidence bin for Experiment 1.

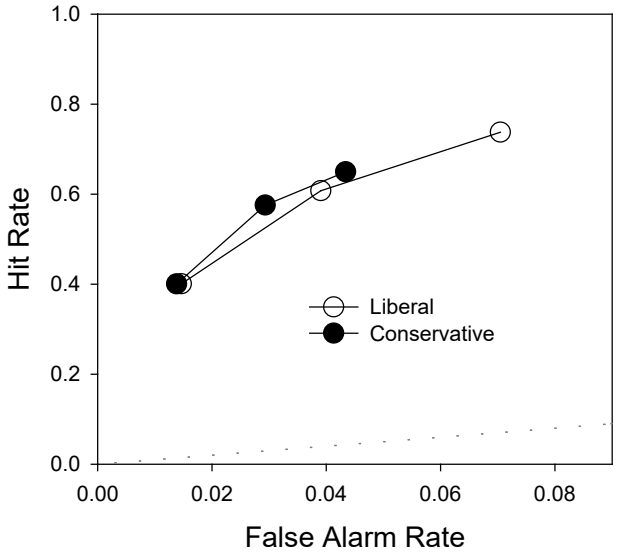
		Liberal			Conservative		
IDs	**Conf**	**TP (S)**	**TP (F)**	**TA**	**TP (S)**	**TP (F)**	**TA**
Positive	High	225	14	54	238	15	48
	Med	116	20	90	104	23	54
	Low	73	45	116	44	18	49
Negative	High	25		201	38		211
	Med	18		90	47		122
	Low	25		69	67		96

Note. TP(S) = suspect IDs from target-present lineups, TP(F) = filler IDs from target-present lineups, and TA = filler IDs (positive) and lineup rejections (negative) from target-absent lineups.

Figure 3 presents the receiver operating characteristic (ROC) data. An ROC is a plot of the hit rate (suspect IDs from TP lineups divided by the number of TP lineups) vs. the false alarm rate (estimated suspect IDs from TA lineups divided by the number of TA lineups) for three different decision criteria. Because there was no designated innocent suspect, the false ID rates were estimated by dividing the TA filler ID rates by lineup size (6). The left most point for each condition only counts suspect IDs made with high confidence, the middle point counts suspect IDs made with medium or high confidence, and the rightmost point counts suspect IDs made with low, medium, or high confidence. The rightmost points represent what is ordinarily

considered to be the overall hit and false alarm rates, and it is visually apparent that it falls farther to the right in the liberal condition (reflecting the more liberal response bias). The two curves trace out essentially the same trajectory, indicting similar levels of discriminability (i.e., the response bias manipulation did not have the unintended consequence of differentially affecting discriminability). This is consistent with an earlier study by Mickes et al. (2017), which found that although discriminability was lower for the liberal and conservative conditions relative to an unbiased condition, they were similar to each other.

Figure 3. ROC data from the liberal and conservative conditions of Experiment 1. The dashed line represents chance performance.



The results of primary interest for Experiment 1 (namely, the CAC results) are shown in Figure 4. For lineup rejections, accuracy within a confidence bin was computed using this formula: nTA / (nTA + nTP), where nTA is the number of target-absent lineup rejections made with a given level of confidence, and nTP is the number of target-present lineup rejections made with a given level of confidence. For positive suspect IDs, accuracy within a confidence bin was computed using this formula: $nTP_{Suspect}$ / ($nTA_{Suspect}$ + $nTP_{Suspect}$), where $nTA_{Suspect}$ is the

number of target-absent suspect IDs made with a given level of confidence, and $nTP_{Suspect}$ is the number of target-absent suspect IDs made with a given level of confidence. Note that, as is typical, $nTA_{Suspect}$ was estimated by dividing the number of filler IDs from TA lineups by lineup size (6).

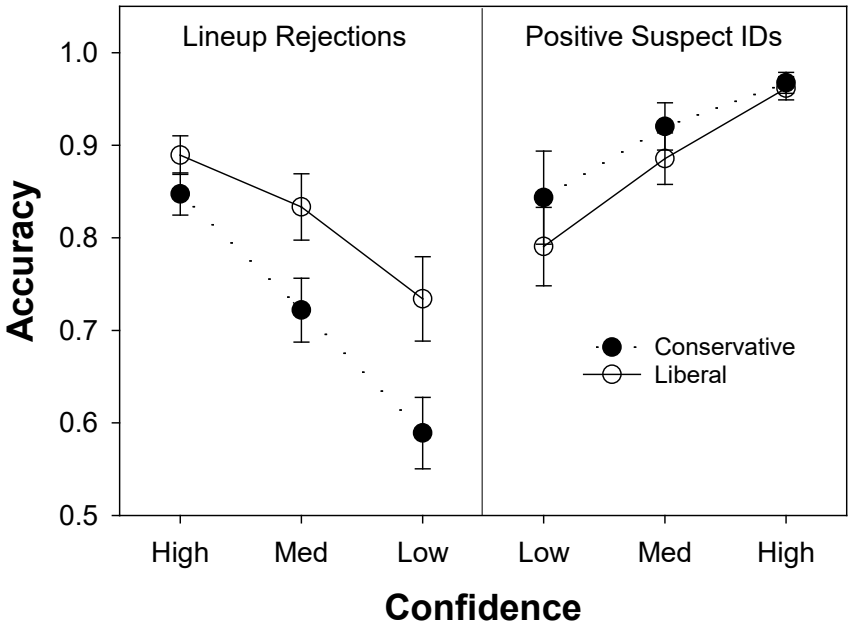


Figure 4. (Left panel) Confidence-accuracy characteristic for lineup rejections in the conservative and liberal conditions. (Right panel) Confidence-accuracy characteristic for positive IDs in the conservative and liberal conditions. The scale on the x-axis can be conceptualized as a 6-point confidence scale, where 1 means "I am sure the perpetrator is not in the lineup" and 6 means "I am sure this person is the perpetrator."

Overall, the effect of the response bias manipulation yielded fairly small effects, but they were in the predicted direction. That is, collapsed over confidence, accuracy for positive suspect IDs in the conservative condition (93.9% correct) was somewhat higher than accuracy in the liberal condition (90.5% correct), $\chi^2(1) = 3.32, p = .069$. At the same time, accuracy for lineup rejections in the conservative condition (73.8% correct) was somewhat lower than accuracy in the liberal condition (84.1% correct), $\chi^2(1) = 15.29, p < .001$.

Within each confidence level considered individually (low, medium, high), pairwise comparisons for positive suspect IDs did not differ significantly for the conservative and liberal conditions. For lineup rejections, accuracy within confidence levels was significantly lower in the conservative condition (relative to the liberal condition) for low and medium confidence, $\chi^2(1) = 5.4, p = 0.014$, and $\chi^2(1) = 4.56, p = 0.033$, respectively. By contrast, the difference for high-confidence lineup rejections was not significant, $\chi^2(1) = 1.81, p = 0.178$.

These effects are consistent with a slope difference for lineup rejections, but the most direct test would be to fit straight lines to each function and statistically compare their slopes. The slope of the CAC function for positive suspect IDs was slightly flatter in the conservative condition (0.06) compared to the liberal condition (0.09), and the slope of the CAC function for lineup rejections was slightly steeper in the conservative condition (-0.13) compared to the liberal condition (-0.08). Both of these effects were in the predicted direction, but a bootstrap statistical analysis was not significant in either case (z = 0.87, p = .386, and z = 1.53, p = .126, respectively).

On balance, the results support the idea that the strength of the confidence-accuracy relationship for both positive IDs and lineup rejections is, at least in part, determined by response bias. However, the effects in Experiment 1 were fairly small, so in Experiment 2, we used a different method of manipulating response bias that allowed for a more decisive test.

Experiment 2

In Experiment 2, everything was the same as in Experiment 1 except that we switched to a forced-choice procedure. Now, participants were always asked to choose the one lineup member who was most likely to be the perpetrator from the crime video. In addition, for the identified individual, participants were also asked to rate their confidence on a -100 to +100

scale, where -100 indicated complete certainty that the identified individual was not the perpetrator, and +100 indicated complete certainty that the identified individual was the perpetrator (0 represented complete uncertainty).

An assumption underlying this experiment is that the -100 to +100 confidence scale represents memory strength, with no point on the scale reflecting anything other than an arbitrary demarcation. Thus, for example, the 0-value is the point at which a participant has decided that memory strength is strong enough to make a positive ID. However, a basic tenet of signal detection theory is that there is nothing particularly special about that 0-value (or any other value) on the continuous memory-strength scale. A more liberal setting for making a positive ID (e.g., -50) or more conservative setting (e.g., +50) would be just as valid. Therefore, after collecting these confidence ratings, we were able to effectively manipulate the decision criterion after the face to determine its effect on the confidence-accuracy relationship for positive IDs and lineup rejections.

Participants

We recruited participants from Amazon's MTurk (n = 3,023). We excluded 106 people due to having seen the stimulus video before. This left 2,917 participants in the final analysis. Participants were compensated 25 or 50 cents for their time. The participants included 41.68% Male (1,214), 57.08% Female (1,665), 0.65% Other (19), 0.51% Decline to Answer (15), and 0.14% no response (4). The ethnicity distribution of the participants was: 76.69% Caucasian (2,237), 8.98% African-American (262), 5.93% Asian (173), 5.07% Latino (148), 0.48% Native-American (14), 0.21% Middle-Eastern (6), 0.14% Pacific-Islander (4), 1.44% Other (42), 0.51% Decline to Answer (15), and 0.55% no response (16).

Design and Materials

The study was a randomized 2 (standard simultaneous vs. 6AFC simultaneous) x 2 (target present vs. target absent) design. The experiment used the same mock-crime stimulus video and 45-second distractor task as above.

Procedure

After viewing the mock-crime video and completing the distractor task, participants moved to the lineup phase. The first set of instructions for the lineup phase of Experiment 2 read as follows: “Imagine you are participating in a real police investigation, and the video you watched showed a real perpetrator committing a real crime. On the next page, you will be presented with some photos (also known as a "lineup"). The lineup may or may not contain the perpetrator of the crime you witnessed. On the next screen, you will receive important instructions along with the lineup. Please follow these instructions carefully.”

After clicking the “Next” button, participants then received one of two lineup conditions, either for a standard simultaneous lineup or a 6AFC simultaneous lineup. The standard simultaneous lineup had two rows of three photographs. In the target present condition, one photo in the lineup was of the guilty suspect while the other five photographs were fillers. The target absent condition did not contain a photo of the perpetrator. Instead, the lineup included a sixth filler photo. Filler photos in the lineup were randomly selected from a pool of 60 possible fillers, all description-matched to the guilty suspect. Participants could select a photograph as being the man from the video or they could reject the lineup by indicating that the man from the video was not present. At the top of the lineup, an instruction read, “Below is a lineup that *may or may not* contain the perpetrator from the video. If you believe that the perpetrator is present, please select his face. Otherwise, please click ‘Not Present’ below.” After participants selected a

face or rejected the lineup, they give their confidence (1-100; 1 = completely unsure; 100 = completely sure).

In the 6AFC condition, participants were shown a lineup with two rows of three photos, with target present and target absent lineups constructed in the same manner as the standard condition. However, for the 6AFC procedure, the instructions at the top of the lineup read: "Below is a lineup that may or may not contain the perpetrator from the video. At the bottom of the lineup, please indicate how sure you are that the perpetrator is or is not in the lineup." Participant would give their confidence (-100 = Completely sure that the man from the video is **not** present in the lineup; 0 = Completely *unsure* whether the man from the video is present in the lineup; +100 = Completely sure that the man from the video **is** present in the lineup). After they answered this detection question and submitted their confidence, they received a new instruction for the same lineup with the same photographs in the same position. The new instructions read, "Note: You are viewing the same lineup as on the last page. If you *had* to choose someone from the lineup as being the perpetrator: 1) Who would you choose and 2) How confident are you that the person is or is not the perpetrator? Please select a face by clicking on it, then indicate your confidence below." After they selected a face, they issued their confidence (-100 = Completely sure that it is **not** the man from the video; 0 = Completely *unsure* whether it is the man from the video; +100 = Completely sure that it **is** the man from the video).

Although we gathered confidence twice in this experiment, (once through a detection question and once through a 6AFC procedure), the ratings ended up being redundant, almost exclusively (i.e., the first and second ratings were almost always the same). Thus, we analyzed the confidence corresponding to the 6AFC question, varying the effective location of the decision criterion.

Although we demarcated a 0-value as being "completely unsure" for both questions, the decision criterion theoretically could exist anywhere within this range as the values are monotonically ordered. We analyzed the 6AFC condition using five different values as the decision criterion (+80, +50, 0, -50, and -80). The criteria of +80 and +50 reflected a more conservative response bias for positive IDs. A criterion of 0 reflected a neutral response bias. The criteria of -50 and -80 reflected a liberal response bias for positive IDs. A positive ID was counted as any confidence value that exceeded that decision criterion, while a confidence value that did not pass that criterion was counted as a lineup rejection. For the standard condition, there was no manipulation of response bias. Positive and lineup rejections were determined by whether the participant selected a face or chose to click the "Not Present" button.

Results

For the standard condition, the bins for low, medium, and high confidence were constructed in the same way as in Experiment 1 (i.e., for positive IDs: 100 to 90 = high confidence; +89 to +70 = medium confidence; +69 to +1 = low confidence for positive IDs; for lineup rejections: -100 to -90 = high confidence; -89 to -70 = medium confidence; -69 to -1 = low confidence). For the 6AFC (neutral response bias) condition, slightly different values were used (namely, +100 to +70 = high confidence; +69 to +25 = medium confidence; +24 to 0 = low confidence for positive IDs and -100 to -82 = high confidence; -81 to -43 = medium confidence; -42 to -1 = low confidence for lineup rejections). In both cases, this scheme was adopted to achieve a relatively large number of ratings falling within each bin so that accuracy scores could be computed with some degree of precision. Table 2 shows the frequency counts for each confidence bin.

Table 2. Frequency counts for each confidence bin in Experiment 2.

		Standard Lineup			6AFC (Neutral)		
IDs	Conf	TP (S)	TP (F)	TA	TP (S)	TP (F)	TA
Positive	High	147	6	19	362	75	121
	Med	184	14	57	150	44	125
	Low	184	46	149	93	43	118
Negative	High	21		148	32	17	151
	Med	50		194	32	20	129
	Low	85		192	32	14	76

Note. TP(S) = suspect IDs from target-present lineups, TP(F) = filler IDs from target-present lineups, and TA = filler IDs (positive) and lineup rejections (negative) from target-absent lineups.

Figure 5 presents the ROC data for the two conditions of Experiment 2. For the Standard condition, the points represent positive suspect IDs. As is typical of lineup ROC data, it is not possible to plot suspect ID (hit) rates for lineup rejections because no face is identified when a lineup is rejected. For the 6AFC condition, by contrast, participants identified the MAX face and supplied a confidence rating even when the lineup was rejected. ROC points for lineup rejections could therefore be plotted even for rejections. That is, for TP lineups, it was known when the MAX rejected face was the guilty suspect (making it possible to plot the "hit rate" even when the face was technically rejected) and for TA lineups, the innocent suspect would be the identified MAX face 1/6 of the time. The ROC points for positive IDs and lineup rejections for the 6AFC condition are connected by a dotted line to create one continuous ROC curve. As in Experiment 1, the two curves trace out essentially the same trajectory, indicting similar levels of discriminability (i.e., the 6AFC requirement did not have the unintended consequence of affecting discriminability relative to the standard condition). Instead, for positive IDs (the leftmost 3 points), the 6AFC condition resulted in a more liberal response bias. The effect was not problematic because our focus was on the slope of the CAC curves as response bias varied over a wide range.

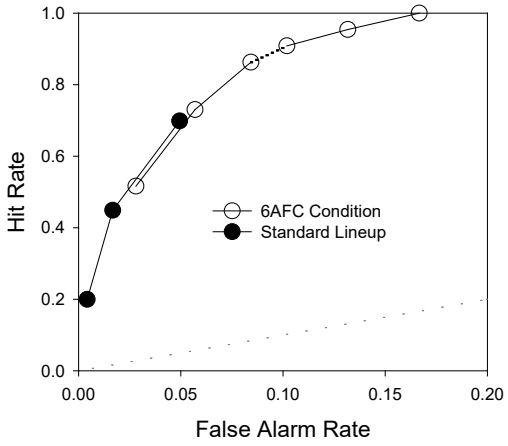
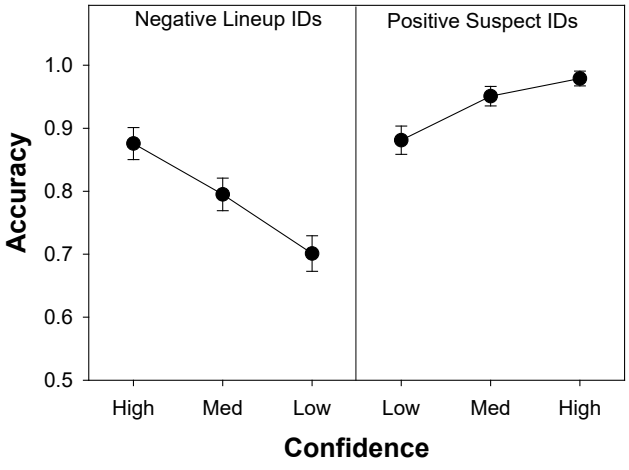


Figure 5. ROC data from the Standard Lineup and 6AFC Condition of Experiment 2. For the 6AFC condition, the leftmost three ROC points (open circles) represent positive IDs (as do the filled circles for the standard condition), whereas the rightmost three ROC points connected by a dotted line represent lineup rejections. The dashed diagonal line represents chance performance.

The CAC results for positive and lineup rejections from the standard condition are shown in Figure 6.[1] Interestingly, and contrary to what is typically observed, the confidence-accuracy relationship is somewhat stronger for lineup rejections than for positive IDs. As described next, this pattern likely reflects the fact that, for whatever reason, the participants in the standard lineup condition of this experiment exhibited a fairly conservative response bias.



[1] The innocent suspect ID rate for this analysis was again estimated by dividing the number of filler IDs from target-absent lineups by lineup size (6).

Figure 6. (Left panel). Confidence-accuracy characteristic for lineup rejections in the standard lineup condition of Experiment 2. (Right panel). Confidence-accuracy characteristic for positive IDs in the standard lineup condition of Experiment 2.

For the 6AFC procedure, for analytical purposes, the location of the decision criterion (nominally set at 0 on the confidence scale) was varied from liberal to conservative. In particular, we set the effective decision criterion to -80, then to -50, then to 0, then to +50, and finally to +80. As an example, with the decision criterion set to -50, any rating above that value was classified as a positive identification of the person who was selected from the lineup as one most likely to be the perpetrator. The binning for classifying such ratings as high, medium, or low confidence changed based on the position of the decision criterion, with the bins chosen to equate the number of observations in each bin as much as possible.

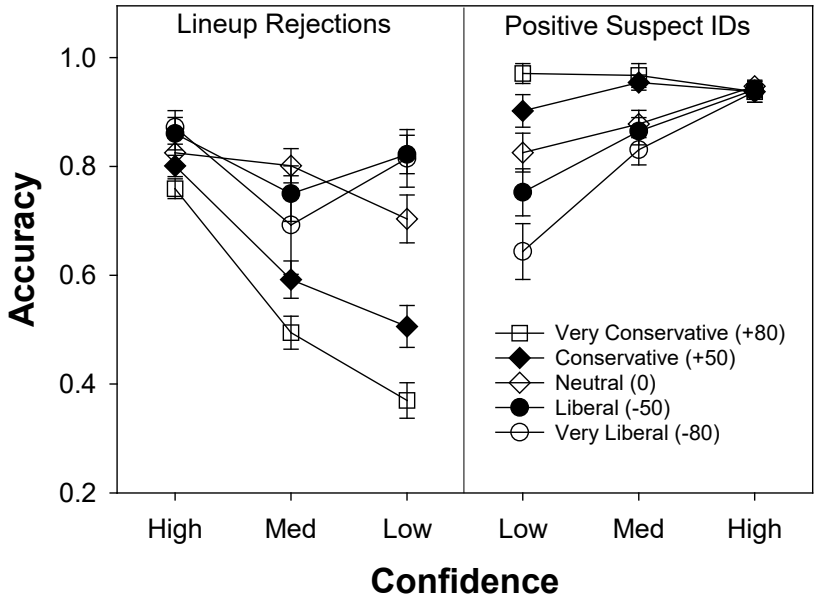


Figure 7. (Left panel). Confidence-accuracy characteristic for lineup rejections in the 6AFC condition of Experiment 2. (Right panel). Confidence-accuracy characteristic for positive IDs in the 6AFC condition of Experiment 2. The scale on the x-axis can be conceptualized as a 6-point confidence scale, where 1 means "I am sure this person is not the perpetrator" and 6 means "I am sure this person is the perpetrator."

As shown in the right panel of Figure 7, the slopes for positive IDs for each decision criterion condition were ordered as predicted. That is, the slope of the confidence-accuracy

relationship was steepest in the most liberal condition (-80) and shallowest in the most conservative condition (+80). Indeed, across all response bias conditions, the slopes were monotonically ordered (they became shallower as the response bias became more conservative).

As shown in the left panel of Figure 7, for lineup rejections, the pattern is somewhat noisier. However, as predicted, the trends are the opposite of the trends observed for positive IDs. For lineup rejections, the confidence-accuracy relationship is the strongest (i.e., the slope is the steepest) for the most conservative condition (+80). The relationship is still strong but is slightly weaker for the conservative (+50) condition, and it is weaker still for the neutral (0) condition. For the two most liberal conditions (-50 and -80), the confidence-accuracy relationship is largely flat for the two endpoints (low vs. high confidence) but dips to a lower value for medium confidence. However, these intermediate medium-confidence points were computed from few observations (18 and 8, respectively). Thus, we assume the dip is due to noise and therefore broke down the confidence for lineup rejections into two confidence bins instead of three for analytical purposes. To re-compute the CACs for lineup rejections using a two-point confidence scale (high vs. low), we distributed the medium-confidence values into the low- and high-confidence bins such that the frequency counts for each confidence level remained roughly equated. We then computed the two-point slopes for each response bias condition. For four out of the five response bias conditions, these two-point slopes for lineup rejections were ordered as predicted (whereas all five of the two-point slopes for positive IDs were ordered as predicted).

To determine how often this pattern of results for positive and lineup rejections would arise by chance, we computed a statistic consisting of the sum of squared differences between the predicted and observed rankings of slopes. For example, if the predicted order across the five

conditions was 1, 2, 3, 4, 5, and if the observed order was 1, 2, 3, 5, 4 (the last two reversed relative to predictions, as in the lineup rejection data here), the statistic would be $(1-1)^2 + (2-2)^2 + (3-3)^2 + (4-5)^2 + (5-4)^2 = 2$. Next, we ran 10,000 bootstrap trials in which the observed rank order was randomly determined. For example, if the random order on a given bootstrap trial was 3, 1, 4, 5, 2, the bootstrap statistic for this trial would be $(1-3)^2 + (2-1)^2 + (3-4)^2 + (4-5)^2 + (5-2)^2 = 13$. We asked how often these randomly ordered bootstrap trials yielded a sum of squares statistic as small or smaller than the observed sum of squares statistics for positive IDs and lineup rejections separately. The result was significant for both positive IDs ($p = 0.008$) and lineup rejections ($p = 0.040$).

General Discussion

The experiments reported here investigated the asymmetrical relationship between confidence and accuracy for positive suspect IDs vs. lineup rejections. Much prior research found a strong confidence-accuracy relationship for positive IDs while simultaneously finding a much weaker relationship for lineup rejections. Yet not all studies show this pattern. Sometimes, the confidence-accuracy relationship for positive IDs is weak (as it was here for the standard lineup condition in Experiment 2), and sometimes, the confidence-accuracy relationship for lineup rejections is fairly strong (e.g., Yilmaz et al., 2022). What explains the usual asymmetry that is observed and the variability that is also sometimes observed across studies?

Here, we propose that differences in response bias provide at least part of the explanation. Using a signal-detection framework (Figure 1), we predicted that a more liberal response bias for positive IDs would yield to a large range of possible values for positive IDs, leading to a strong confidence-accuracy relationship. At the same time, it would yield a smaller range of possible values for lineup rejections—thereby leading to a flatter confidence-accuracy function for lineup

rejections. A more conservative response bias for positive IDs would have the opposite effect, weakening the confidence-accuracy relationship for positive IDs and strengthening it for lineup rejections.

To test these predictions, in Experiment 1, we manipulated response bias using lineup instructions designed to elicit conservative or liberal responding. The hypothesis was that liberal response bias for making positive IDs would yield a strong confidence-accuracy relationship for positive IDs and a weaker confidence-accuracy relationship for lineup rejections. Conversely, we predicted that a conservative response bias for positive IDs would yield to a weaker confidence-accuracy relationship for positive IDs and a stronger confidence-accuracy relationship for lineup rejections. Though the effects were small, the results for Experiment 1 turned out as predicted.

Experiment 2 used a 6AFC procedure that allowed us to manipulate response bias more effectively (after the fact) based on the confidence ratings provided by the participants. The results were again largely (and more convincingly) in accordance with our predictions. That is, the steepness of the slope (i.e., the strength of the relationship between confidence and accuracy) for positive and lineup rejections varied in opposite directions as a function of response bias.

Two other factors, not investigated here, might also affect the strength of the confidence-accuracy relationship for lineup rejections. One factor is whether the decision variable itself might be causing the asymmetric empirical pattern of data shown earlier in Figure 2. Functionally for lineups, confidence for positive IDs is given in relation to a single face (i.e., the selected face, with the MAX memory signal). However, it is less clear what confidence is tied to for lineup rejections since the task for simultaneous lineups involves collectively rejecting a set of faces. Conceivably, confidence in lineup rejections is based on the average memory signal rather than on the MAX memory signal (as posited by Brewer & Wells, 2006; Lindsay et al.,

2013; Yilmaz et al., 2022). The use of an average signal might yield a weaker confidence-accuracy relationship. However, recent research from our lab suggests this explanation may not be right. Using a model-fitting approach, we found evidence supporting the idea that confidence is based on the MAX face regardless of whether a positive ID or a lineup rejection is made (Yilmaz & Wixted, in press).

A second factor that may indirectly influence the strength of the confidence-accuracy relationship for lineup rejections is the overall level of performance on the lineup task. When performance is very high, as it was in the simultaneous condition of Experiment 2, participants might choose to adopt a conservative decision criterion such that accuracy is high whether confidence is low or high (i.e., the confidence-accuracy relationship for positive IDs would be weak). If so, one would expect to see a stronger confidence-accuracy relationship for lineup rejections, as we did here for the standard lineup condition in Experiment 2. The opposite would be true when overall performance is worse. Whether this factor might also help to explain the mystery of the (typically) weak confidence-accuracy relationship for lineup rejections remains to be seen.

Whatever the explanation turns out to be, achieving a better understanding of the relationship between confidence and accuracy for lineup rejections seems important given that many of the DNA exoneration cases involving high-confidence misidentifications at trial began with something other than that (sometimes with a lineup rejection) on the initial test (Garrett, 2011). It is essential to focus on the results of the first test (Wells et al., 2020; Wixted et al., 2021), especially when the witness rejects the lineup, but a key question that has not yet been fully answered is when confidence informs accuracy for lineup rejections. The main finding reported here is that confidence in a lineup rejection is more informative when response bias is

conservative compared to when it is liberal. Thus, if these results are confirmed by other labs using different stimulus materials, then for jurisdictions that use lineup instructions to encourage a conservative response bias, it would be safe to conclude that confidence in a lineup rejection has more information value than would otherwise be the case.

References

Arndorfer, A., & Charman, S. D. (2022). Assessing the effect of eyewitness identification confidence assessment method on the confidence-accuracy relationship. *Psychology, Public Policy, and Law, 28*(3), 414–432. https://doi.org/10.1037/law0000348

Brewer, N., Keast, A., & Rishworth, A. (2002). The confidence-accuracy relationship in eyewitness identification: The effects of reflection and disconfirmation on correlation and calibration. *Journal of Experimental Psychology: Applied, 8*(1), 44–56. https://doi.org/10.1037/1076-898X.8.1.44

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

Carlson, C. A., Dias, J. L., Weatherford, D. R., & Carlson, M. A. (2017). An investigation of the weapon focus effect and the confidence–accuracy relationship for eyewitness identification. *Journal of Applied Research in Memory and Cognition, 6(1)*, 82–92. https://doi.org/10.1037/h0101806

Clark, S. E. (2005). A Re-examination of the Effects of Biased Lineup Instructions in Eyewitness Identification. *Law and Human Behavior, 29*(4), 395–424. https://doi.org/10.1007/s10979-005-5690-7

Dobolyi, D. G., & Dodson, C. S. (2013). Eyewitness confidence in simultaneous and sequential lineups: A criterion shift account for sequential mistaken identification overconfidence. *Journal of Experimental Psychology: Applied*, *19*(4), 345–357. https://doi.org/10.1037/a0034596

Dodson, C. S., & Dobolyi, D. G. (2016). Confidence and eyewitness identifications: The cross-race effect, decision time and accuracy. *Applied Cognitive Psychology, 30*(1), 113–125. https://doi.org/10.1002/acp.3178

Garrett, B. (2011). *Convicting the innocent: Where criminal prosecutions go wrong*. Cambridge, MA: Harvard University Press.

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*(4), 346–360. https://doi.org/10.1037/a0029779

Juslin, P., Olsson, N., & Winman, A. (1996). Calibration and diagnosticity of confidence in eyewitness identification: Comments on what can be inferred from the low confidence–accuracy correlation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*(5), 1304–1316. https://doi.org/10.1037/0278-7393.22.5.1304

Keast, A., Brewer, N., & Wells, G. L. (2007). Children's metacognitive judgments in an eyewitness identification task. *Journal of Experimental Child Psychology*, *97*(4), 286–314. https://doi.org/10.1016/j.jecp.2007.01.007

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

Mickes, L., Seale-Carlisle, T. M., Wetmore, S. A., Gronlund, S. D., Clark, S. E., Carlson, C. A., Goodsell, C. A., Weatherford, D., & Wixted, J. T. (2017). ROCs in eyewitness

identification: Instructions versus confidence ratings. *Applied Cognitive Psychology, 31*(5), 467–477. https://doi.org/10.1002/acp.3344

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*(1), 55–71. https://doi.org/10.1037/a0031602

Quigley-McBride, A., & Wells, G. L. (2023). Eyewitness confidence and decision time reflect identification accuracy in actual police lineups. *Law and Human Behavior, 47*(2), 333–347. https://doi.org/10.1037/lhb0000518

Sauer, J. D., Brewer, N., & Wells, G. L. (2008). Is there a magical time boundary for diagnosing eyewitness identification accuracy in sequential line-ups? *Legal and Criminological Psychology*, *13*(1), 123–135. https://doi.org/10.1348/135532506x159203

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*(4), 337–347. https://doi.org/10.1007/s10979-009-9192-x

Sauerland, M., & Sporer, S. L. (2009). Fast and confident: Postdicting eyewitness identification accuracy in a field study. *Journal of Experimental Psychology: Applied*, *15*(1), 46–62. https://doi.org/10.1037/a0014560

Smith, A. M., Ayala, N. T., & Ying, R. C. (2023). The rule out procedure: A signal-detection-informed approach to the collection of eyewitness identification evidence. Psychology, Public Policy, and Law, 29(1), 19–31. https://doi.org/10.1037/law0000373

Technical Working Group for Eyewitness Evidence. (1999). Eyewitness evidence: A guide for law enforcement. Washington, DC: U.S. Department of Justice, Office of Justice Programs.

Weber, N., & Brewer, N. (2004). Confidence-Accuracy Calibration in Absolute and Relative Face Recognition Judgments. *Journal of Experimental Psychology: Applied*, *10*(3), 156–172. https://doi.org/10.1037/1076-898x.10.3.156

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

Wixted, J. T., Mickes, L., Clark, S. E., Gronlund, S. D., & Roediger, H. L., 3rd (2015). Initial eyewitness confidence reliably predicts eyewitness identification accuracy. *The American psychologist*, *70*(6), 515–526. https://doi.org/10.1037/a0039510

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, 10-65.

Wixted, J. T., Wells, G. L., Loftus, E. F., & Garrett, B. L. (2021). Test a witness's memory for a suspect only once. *Psychological Science in the Public Interest, 22(1_suppl)*, 1S-18S. https://doi.org/10.1177/15291006211026259

Yilmaz, A.S., & Wixted, J.T. (submitted). What Latent Variable Underlies Confidence in Lineup Rejections?

Yilmaz, A. S., Lebensfeld, T. C., & Wilson, B. M. (2022). The reveal procedure: A way to enhance evidence of innocence from police lineups. *Law and Human Behavior*, *46*(2), 164–173. "https://doi.org/10.1037/lhb0000478"https://doi.org/10.1037/lhb0000478