Simultaneous Versus Sequential Presentation in Testing Recognition Memory for Faces

JASON R. FINLEY, HENRY L. ROEDIGER III, ANDREA D. HUGHES, CHRISTOPHER N. WAHLHEIM and LARRY L. JACOBY
Washington University in St. Louis

Three experiments examined the issue of whether faces could be better recognized in a simultaneous test format (2-alternative forced choice [2AFC]) or a sequential test format (yes–no). All experiments showed that when target faces were present in the test, the simultaneous procedure led to superior performance (area under the ROC curve), whether lures were high or low in similarity to the targets. However, when a target-absent condition was used in which no lures resembled the targets but the lures were similar to each other, the simultaneous procedure yielded higher false alarm rates (Experiments 2 and 3) and worse overall performance (Experiment 3). This pattern persisted even when we excluded responses that participants opted to withhold rather than volunteer. We conclude that for the basic recognition procedures used in these experiments, simultaneous presentation of alternatives (2AFC) generally leads to better discriminability than does sequential presentation (yes–no) when a target is among the alternatives. However, our results also show that the opposite can occur when there is no target among the alternatives. An important future step is to see whether these patterns extend to more realistic eyewitness lineup procedures.

The pictures used in the experiment are available online at http://www.press.uillinois.edu/journals/ajp/media/testing_recognition/.

Recognition tests are perhaps the most studied procedure in experimental research on memory processes. A typical procedure might begin with participants studying 100 items (e.g., words, pictures, or faces) and then being tested on 200 items: 100 studied items (or targets) and 100 nonstudied items (or lures). Two standard forms of recognition test have been developed: free choice (or yes–no) and forced choice. In a free-choice test, each trial shows one of the 200 test items, and participants judge whether that item was studied (old, yes) or nonstudied (new, no); participants are free to designate any number of items as being old, hence the name *free choice*. In a two-alternative forced-choice test (2AFC), each trial shows two items together, one target and one lure, and participants are forced to choose which of the two was studied. These two basic procedures can be manipulated in many ways, and other judgments, such as confidence ratings, can be added.

Psychologists interested in basic memory processes have used one procedure or the other for many purposes and have rarely been concerned about possible differences between them. Signal detection theory (SDT) is a framework that has been widely used in the study of recognition memory, and it holds that yes–no and forced-choice procedures produce similar outcomes in terms of discriminability of targets from lures. In fact, SDT predicts equal

American Journal of Psychology
Summer 2015, Vol. 128, No. 2 pp. 173–195 • © 2015 by the Board of Trustees of the University of Illinois

values of d' under the two procedures if the appropriate adjustment is applied (dividing by $\sqrt{2}$ in the case of 2AFC; Macmillan & Creelman, 2005, p. 168, equation 7.2; see also McNicol, 2004, pp. 175–176). The adjustment depends on several assumptions, including that judgments are made independently for members of a forced-choice pair, meaning that the basis for judgments is the same for forced-choice and single-item tests (Green & Swets, 1966, pp. 48, 68). Some evidence supporting the assumptions of SDT has been obtained in recognition memory research (Green & Moses, 1966; Jang, Wixted, & Huber, 2009). However, there has also been evidence against the equivalence of discriminability in yes–no versus 2AFC procedures (Deffenbacher, Leu, & Brown, 1981; Kroll, Yonelinas, Dobbins, & Frederick, 2002).

The situation is very different in one applied setting that uses a modified version of the standard free- and forced-choice tests: eyewitness recognition of suspected criminals in lineups (either photographic or in person). In both types of lineup, a suspect appears with other people of the same general description (often five), and an eyewitness is asked to select the perpetrator from the set, if he is in fact present (see Wells et al., 1998, for an overview of the issues involved in constructing lineups). In a simultaneous lineup, all six candidates appear at once. This simultaneous presentation procedure is similar to a forced-choice recognition procedure with the exceptions that the real perpetrator may or may not be present, and the eyewitness may choose not to select anyone (or to "reject the lineup"). The hope from any lineup situation is that it will maximize identification of guilty suspects (hits) and minimize erroneous identification of innocent people (false alarms). However, it has been known for many years that eyewitness identification is far from perfect, and consequently a substantial number of innocent people have been convicted almost entirely on the basis of false identification (Buckhout, 1974; Garrett, 2011; see also The Innocence Project at http://www.innocenceproject.org).

The effectiveness of the standard simultaneous lineup procedure was compared with another procedure, the sequential lineup, in a landmark study by Lindsay and Wells (1985). A sequential lineup differs from a simultaneous lineup in that an eyewitness views potential perpetrators individually and is instructed to make a recognition memory decision about each person. The lineup ends either when the eyewitness identifies a perpetrator or when no identification has been made after the viewing of all potential perpetrators. This procedure resembles a standard yes–no recognition test. Lindsay and Wells's results were dramatic in showing that although there was no significant difference in the proportion of correct identifications (hit rate) between the sequential and simultaneous procedures (.50 vs. .58), the proportion of false identifications (false alarm rate) was substantially lower for sequential versus simultaneous (.17 vs. .43). They argued that a simultaneous lineup encourages eyewitnesses to use a relative judgment process to select the candidate who looks most like the perpetrator they remember seeing (Wells, 1984) and that this process tends to yield false identifications when the true perpetrator is not in the lineup. In contrast, sequential lineups encourage an absolute judgment process in which eyewitnesses individually compare each candidate with their memory of the perpetrator, and this process is less likely to lead to false identification.

Subsequent research has generally confirmed that sequential lineups produce lower false alarm rates than simultaneous lineups, but not by the large magnitude obtained by Lindsay and Wells (1985). A recent meta-analysis of relevant experiments shows that both the hit rate and the false alarm rate are lower in the sequential lineup procedure (Steblay, Dysart, & Wells, 2011). This outcome may indicate that sequential lineups simply induce more conservative responding; eyewitnesses are less likely to give a "yes" response at all in the sequential versus the simultaneous lineup. This outcome would then still leave open to debate the issue of which lineup provides better discriminability, and thus which is to be generally preferred for forensic purposes.

The experiments reported in this manuscript were designed to help answer this question in a simplified face recognition situation. This set of experiments was begun in 2005 by two of the authors (L.L.J. and A.D.H.), but events intervened to delay publication, and in the meantime several other research teams took up this same issue. We describe the rationale behind our research, which we still view as highly pertinent, and then describe recent advances in the field before reporting our experiments.

In most eyewitness memory studies, participants view one simulated crime and later complete a memory test for one lineup. Such procedures yield one datum per participant and thus require large sample sizes. However, more traditional laboratory experiments on face recognition yield numerous observations per participant and can shed light on the underlying cognitive processes relevant to eyewitness identification tasks. We used basic laboratory face recognition tasks and compared free-choice (analogous to a sequential lineup) and forced-choice (analogous to a simultaneous lineup) recognition tests. The analogy is not perfect, but if the results are consistent with other findings from simulated lineup situations, then the outcomes here would gain credence, and the ancillary analyses permitted in our experiments may shed useful light on the issues at hand.

One factor that is relevant to eyewitness identification but has not been fully explored in basic face recognition research is the option for a eyewitness to say, "I don't know." Koriat and Goldsmith (1996) developed a procedure that permits examination of response withholding and how it can affect recognition accuracy. In a first phase, participants took a general knowledge test, either in recall format or multiple-choice format, and made confidence ratings for each response. They were required to answer every question, even if they had to guess. In a second phase, participants took the same test again but this time were allowed to decide whether or not they wanted to provide any answer to each question, and they were offered one of several levels of monetary reward for correct responses and penalty for incorrect ones. Their results showed that when participants were sufficiently monetarily motivated and their metacognitive monitoring was effective, allowing participants to strategically withhold responses increased the proportion of their volunteered responses that were accurate (what Koriat & Goldsmith call output-bound scoring). We used a variant of this free report procedure in our Experiments 2 and 3 to see whether the yes–no and 2AFC test formats enabled equivalent improvements in performance.

The purpose of the present experiments was to investigate whether the yes–no test format or the 2AFC test format yields superior face recognition performance (Experiment 1), whether any such advantages persist when participants are allowed to withhold responses (Experiment 2), and whether any such advantages extend to conditions analogous to the sorts of target-absent procedures used in lineup experiments in which participants are permitted to say that neither response is correct (Experiment 3). Before getting to our experiments, however, we need to provide some recent history of the controversy over comparisons between simultaneous and sequential lineups.

Based on the results of Lindsay and Wells (1985) and other results, Wells (2014) has argued that the best way to assess which lineup procedure is superior is to use a diagnosticity ratio, which is simply the hit rate divided by the false alarm rate. Thus for the Lindsay and Wells results, the diagnosticity ratio for the sequential lineup is .50/.17 = 2.94 and for the simultaneous lineup is .58/.43 = 1.35. The higher the ratio, the argument goes, the better the procedure, because hits (correct identifications) more greatly outweigh false alarms (erroneous identifications). Given these results, some psychologists began strongly recommending that police departments across the United States replace the traditional simultaneous lineup with sequential lineups (Wells et al., 1998), and many departments have done so (Gronlund, Wixted, & Mickes, 2014).

Recently, however, Mickes, Flowe, and Wixted (2012) have challenged this advice based on diagnosticity ratios. They demonstrated that the increase in the diagnosticity ratio results simply from the fact that sequential lineups generally lead to a more conservative response bias relative to simultaneous lineups, with fewer hits and false alarms. To take the case to the extreme, if 1,000 lineups were conducted and 3 led to correct identifications and 1 led to an incorrect identification, the diagnosticity ratio would be greater than ever seen in the literature (3.0), but the procedure would for all practical purposes be worthless; responding is so conservative that in 996 out of 1000 cases, an eyewitness failed to pick any suspect. Mickes et al. (2012; see also Gronlund, Wixted, & Mickes, 2014) thus argued that the diagnosticity ratio is not indicative of the discriminability between guilty and innocent suspects that is yielded by a given procedure and that the measure is thus not relevant for policy decisions (but see Wells, 2014, for a dissent).

Does the sequential lineup procedure in fact lead to greater discriminability than the simultane-

ous procedure, validating the shift in police policy to the sequential lineup? Mickes et al. (2012) compared receiver operator characteristic (ROC) curves for simultaneous and sequential lineup procedures in three eyewitness memory experiments. Participants watched a video of a simulated crime and then responded to a lineup that was either simultaneous or sequential, in which the perpetrator (target) was either present or absent. Participants rated their confidence in their responses using a 100-point scale. This procedure permitted the researchers to produce ROC curves and to compare the two procedures on discriminability (using area under the ROC curve, a measure of memory performance). In Experiment 1a they found that greater discriminability yielded by simultaneous relative to sequential lineups. Experiment 1b, a replication, revealed a similar pattern, although no statistically significant difference appeared. Still, there was certainly no sequential lineup advantage. In Experiment 2 the researchers used a biased lineup for their target-absent condition in which one member of the lineup looked somewhat like the real perpetrator. Again they obtained no difference in discriminability of the lineups, and they concluded that, if anything, the simultaneous lineup produced superior discriminability relative to the sequential lineup.

The Mickes et al. (2012) results are startling because they suggest that psychologists may have been advocating for years that police switch from a lineup procedure that yields superior discriminability (simultaneous) to one that yields inferior discriminability (sequential), and many departments have been following this advice. But can these results be replicated? The answer is yes. Since publication of Mickes et al., at least four other articles have appeared reporting a variety of similar experiments. All the data show either an advantage of simultaneous to sequential lineups in discriminability or no difference between the two. Most of the experiments show an advantage for simultaneous lineups (see Anderson, Carlson, Carlson, & Gronlund, 2014; Carlson & Carlson, 2014; Dobolyi & Dodson, 2013; Gronlund et al., 2012).

The three experiments we report here are pertinent to the debate that has unfolded in the years since they were first conceived. We both replicate the simultaneous testing advantage and also report a reversal: a case in which a sequential recognition procedure (yes–no) produces an advantage in discriminability relative to a simultaneous recognition procedure (2AFC).

EXPERIMENT 1

Participants studied 40 faces and then took either a yes–no or 2AFC recognition test over 80 faces. The lure faces bore either high or low similarity to the target face. Participants made recognition responses and rated their confidence in the responses.

METHOD

Design and Participants

The experiment used a 2 (test format) × 2 (lure similarity) design. Test format (yes–no vs. 2AFC) was manipulated between subjects. Lure similarity to the target (high vs. low) was manipulated within subjects. Dependent measures were recognition responses (yes–no or left–right), and confidence ratings (0–100 for yes–no, 50–100 for 2AFC). Participants were 40 undergraduate students (29 female, mean age 20 years, SD = 1) recruited through the Washington University participant pool who received either $10 or course credit for their participation.[1]

Materials

Materials consisted of 40 sets of three computer-generated face composite pictures created with the software application FACES: The Ultimate Composite Picture (InterQuest Inc., 1998). All faces were male. Each set consisted of an original face, a high-similarity lure, and a low-similarity lure. Figure 1 provides an example. The entire set of faces used is provided in the supplemental materials, available at the first author's Web site (http://jasonfinley.com/) or this journal's Web site (http://www.press.uillinois.edu/journals/ajp/media/testing_recognition/).

The original faces were created using seven features: head shape, jaw, lips, nose, eyes, eyebrows, and hair. The original faces were constructed such that they were not particularly similar to one another. High-similarity lure faces were created by changing the original face on three features: lips, nose, and eyes. Low-similarity faces were created by changing the original face on five features: lips, nose, eyes, eyebrows, and hair. Head shape and jaw were consistent throughout a set. Face pictures were presented on a computer screen at a size of 300 × 400 pixels.

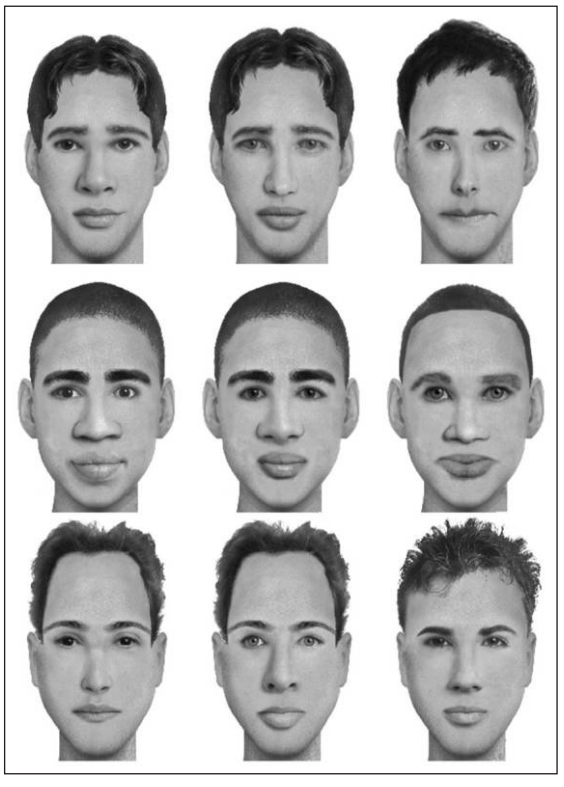

FIGURE 1. Examples of face pictures used in all experiments. Left column: original face. Middle column: high-similarity lure. Right column: low-similarity lure

Procedure

All participants in Experiments 1 and 2 were tested individually on computers using the E-Prime software package (Psychology Software Tools, 2002). Participants in Experiment 1 made their responses using the computer keyboard. The "A" and "L" keys on the keyboard were respectively relabeled as "old" and "new" for the yes–no test or "left" and "right" for the 2AFC test.

The procedure consisted of a study phase followed by a test phase. In the study phase, participants were first instructed that they were about to view a series of faces that they would need to remember for a later memory test. They then viewed the 40 original faces at a rate of 3 s per face with a 500-ms interstimulus interval, in an order that was randomized for each participant. For the test phase, participants were randomly assigned to either the yes–no recognition test format (n = 20) or the two-alternative forced-choice (2AFC) recognition test format (n = 20). There was no intervening task between the study phase and the test phase.

YES–NO RECOGNITION TEST.

On each trial, participants were shown a single face and were instructed to press a key labeled "old" if the face had been presented during the study phase or a key labeled "new" if the face was not presented during the study phase. After responding, they typed in a number to rate their confidence in their answer on a scale of 0% to 100%. There were 80 trials in total, consisting of the 40 originally studied faces and 1 lure corresponding to each studied face (20 high-similarity lures, and 20 low-similarity lures). Face sets were counterbalanced such that across participants each face set was equally represented in the two lure conditions (high similarity vs. low similarity). Test trials occurred in one of four fixed random orders, with the following constraints: Exactly half of the originally studied faces occurred earlier in the test than their corresponding lures (and vice versa), and no lure condition occurred more than three trials in a row. The mean lag between original faces and their corresponding lures was 39.0 intervening trials (SD = 16.6, range 3–76).

2AFC RECOGNITION TEST.

On each trial, participants were shown two faces side by side, and they were instructed to press a key labeled "left" if the face on the left was presented during the study phase or a key labeled "right" if the face on the right was presented during the study phase. After responding, they typed in a number to rate their confidence in their answer on a scale of 50% to 100%. There were 40 trials in total, consisting of 20 trials in which the original face was paired with its high-similarity lure, and 20 trials in which the original face was paired with its low-similarity lure. Face sets were counterbalanced such that across participants each face set was equally represented in the two lure conditions (high similarity vs. low similarity). Test trials occurred in one of four fixed random orders, with the following constraints: In exactly half of the trials, the originally studied face was on the left side (and vice versa), and no lure condition occurred more than three trials in a row.

Note that for both yes–no and 2AFC test formats, two faces (one old and one new) were tested from each face set; in the yes–no test those faces appeared individually on separate trials that were widely spaced on average, and in the 2AFC test those faces appeared together on a single trial.

RESULTS AND DISCUSSION

An $\alpha = .05$ was used for all tests of statistical significance unless otherwise noted. Effect sizes for comparisons of means are reported as Cohen's *d*, calculated using the pooled standard deviation of the groups being compared. Effect sizes for ANOVAs are reported as partial omega-squared, $\hat{\omega}_p^2$, calculated using the formulas provided by Maxwell and Delaney (2004, p. 598). Omega-squared is preferred because it is a less biased estimator of population effect size than eta-squared; the partial formulation gives the variance in the dependent measure accounted for by one particular independent variable, with the effects of other variables in the design partialled out. Standard deviations (*SD*s) are reported raw (i.e., calculated using N, not $N - 1$), on the grounds that the *SD* is a descriptive statistic, and the $N - 1$ adjustment should be reserved for use in inferential statistics.

We pause now to make several clarifications about the performance measures we will be reporting. For 2AFC, we define hit rate as the proportion of trials on which the participant chose the correct face and false alarm rate as the proportion of trials on which the participant chose the incorrect face. In Experiment 1, the 2AFC false alarm rate is simply 1 minus the hit rate, but this will not necessarily be the case in Experiments 2 and 3 (because of the option to volunteer or withhold a response, and the addition of a "neither" option in Experiment 3).

Although we report the standard measures of hit rate, false alarm rate, and d', for all three experiments we will focus our analyses on the measure area under the curve (AUC), which is the area under the ROC curve yielded by signal detection analysis. Empirical ROC curves for each participant were constructed by grouping confidence ratings into 5 bins (i.e., dividing the entire rating scale into fifths) and counting up the cumulative number of hits and false alarms in each successive confidence bin, starting with the highest-confidence bin. This is the ROC construction method described by Mickes et al. (2012, pp. 366–367) for lineup research, and we note that it differs from the ROC construction method traditionally used in basic laboratory recognition research (e.g., Macmillan & Creelman, 2005, pp. 53–57). We calculated AUC using the trapezoidal method of Pollack and Hsieh (1969), who referred to the measure as A^g (see also Macmillan & Creelman, 2005, p. 64). AUC ranges from 0 to 1, with larger values indicating better memory performance (i.e., discriminability). The measure AUC is nonparametric and thus most appropriate for comparing performance between yes–no recognition tests and 2AFC recognition tests. The area under the ROC curve is also what Mickes et al. (2012) recommend for comparing memory performance between simultaneous and sequential lineup procedures.[2]

In addition to reporting AUC we will also report d', which is useful for comparison with other studies where AUC cannot be calculated (see advice from Mickes, Moreland, Clark, & Wixted, 2014). The measure d' is based on signal detection theory, and it summarizes a participant's ability to discriminate studied items from unstudied items, separate from his or her overall bias toward giving one type of response over another (e.g., tendency to say "old" over "new"). As we mentioned earlier, SDT predicts that, all else being equal, d' in 2AFC should be larger than d' in yes–no by a factor of $\sqrt{2}$ (Macmillan & Creelman, 2005, p. 168; see also Kroll et al., 2002, Footnote 1). We will report d' in 2AFC divided by $\sqrt{2}$ only in Experiment 1, in addition to reporting raw d'. We will report only raw d' in Experiments 2 and 3. We have two reasons for this decision. First, in line with our framing of this research as relevant to eyewitness lineup identification, we are interested primarily in the practical matter of which procedure yields better performance, rather than testing the equivalence-with-adjustment that is predicted by SDT. Second, the $\sqrt{2}$ adjustment was developed only for the strict traditional 2AFC procedure, in which one target item and one lure item were present in every trial and in which there was no other option but to choose one of the two items. Our Experiments 2 and 3 include target-absent trials, and Experiment 3 includes a "neither" response option. These procedural differences violate the assumptions under which the $\sqrt{2}$ adjustment applies. Finally, we note that in calculating d' when a hit rate or false alarm rate was equal to 0 or 1, we used the half-point correction method (Macmillan & Creelman, 2005, p. 8).

Recognition Performance

Table 1 shows several measures of recognition performance as a function of test format and lure similarity. The data make two primary points: Recognition

TABLE 1. Mean (*SD*) Performance, Experiment 1

Test format × lure similarity	Hit rate	False alarm rate	*d′*	AUC
Yes–no recognition				
Low similarity	.54 (.12)	.19 (.10)	1.06 (0.49)	.70 (.08)
High similarity	.59 (.10)	.45 (.13)	0.38 (0.41)	.60 (.07)
2AFC recognition				
Low similarity	.78 (.09)	.22 (.09)	1.61 (0.58)	.82 (.08)
High similarity	.70 (.12)	.30 (.12)	1.13 (0.77)	.72 (.12)

Note. d′ *values in table are raw. Using the* $\sqrt{2}$ *adjustment predicted by signal detection theory, mean* d′ *in 2AFC was 1.14 (0.41) for low similarity, and 0.80 (0.55) for high similarity. 2AFC = 2-alternative forced choice; AUC = area under the curve.*

was superior with 2AFC testing relative to yes–no tests and was better when the lures were more dissimilar from the targets. The observations were confirmed with a two-way mixed ANOVA, using AUC as the dependent variable, indicating a main effect of test format, $F(1, 38) = 30.46$, $MSE = .009$, $p < .001$, $\hat{\omega}_p^2 = .425$. Participants were better at discriminating between target and lure faces when the two faces were viewed simultaneously (2AFC) rather than sequentially (yes–no). Not surprisingly, performance was better for items in the low-similarity condition than in the high-similarity condition, $F(1, 38) = 27.20$, $MSE = .007$, $p < .001$, $\hat{\omega}_p^2 = .228$. That is, participants were better at discriminating between target and lure faces when the lure face was less similar to the target face. There was no reliable interaction between test format and lure similarity, $F(1, 38) = 0.02$.

We now briefly report the same analyses using *d′* for the sake of any readers interested in the equivalence with adjustment predicted by SDT. Without any adjustment to *d′* in the 2AFC case, there was a main effect of test format, $F(1, 38) = 19.66$, $MSE = .423$, $p < .001$, $\hat{\omega}_p^2 = .318$, a main effect of lure similarity, $F(1, 38) = 23.55$, $MSE = .286$, $p < .001$, $\hat{\omega}_p^2 = .189$, and no significant interaction, $F(1, 38) = 0.69$, $MSE = .286$, $p = .413$, $\hat{\omega}_p^2 < .001$. When adjusting *d′* by dividing by $\sqrt{2}$ in the 2AFC case, there was a main effect of test format, $F(1, 38) = 4.59$, $MSE = .260$, $p = .039$, $\hat{\omega}_p^2 = .082$, a main effect of lure similarity, $F(1, 38) = 25.81$, $MSE = .201$, $p < .001$, $\hat{\omega}_p^2 = .217$, and no significant interaction, $F(1, 38) = 2.85$, $MSE = .201$, $p = .099$, $\hat{\omega}_p^2 = .020$. Thus, our results are in contrast with the prediction made by SDT and similar to the findings of Deffenbacher et al. (1981) and Kroll et al. (2002). Even when adjusting 2AFC performance (*d′*), participants showed greater discriminability on this test relative to the yes–no test.

In agreement with the lineup experiments cited earlier, the simultaneous viewing of faces seems to improve discrimination relative to yes–no sequential tests. Furthermore, this outcome occurs regardless of whether the lure face is of high or low similarity to the target (although highly similar lures reduce overall discriminability).

Metacognition

To evaluate how well participants' confidence discriminated between correct and incorrect responses, we calculated gamma correlations for each participant as a function of lure similarity (low vs. high) and as a function of item type (target vs. lure) in the case of yes–no recognition.[3] Table 2 shows these results. The correlation between confidence and accuracy was positive in every condition except for high-similarity lures in the yes–no recognition test, where the correlation was reliably negative, $t(19) = 2.38$, $p = .028$. That is, when high-similarity lure faces were viewed in isolation (yes–no), participants were generally more confident in their false alarms than their correct rejections. In an eyewitness testimony setting, such a result could contribute to false convictions.

Negative correlations between confidence and accuracy are surprising but certainly not unprecedented (e.g., Tulving, 1981). This result is consistent with recent work by DeSoto and Roediger (2014; see also Roediger & DeSoto, 2014), who found negative

TABLE 2. Mean (*SD*) Within-Participant Confidence–Accuracy Gamma Correlations, Experiments 1–3

Test format × trial type	Experiment 1		Experiment 2		Experiment 3	
	Low similarity	High similarity	Low similarity	High similarity	Low similarity	High similarity
Yes–no recognition						
Target	.53 (.31)*	.38 (.36)*	.43 (.42)*	.38 (.59)*	.43 (.49)*	.32 (.51)*
Lure (target present)	.25 (.52)	–.19 (.35)*	.29 (.56)*	–.11 (.59)	.26 (.52)*	–.06 (.52)
Lure (target absent)			.25 (.60)*	.16 (.60)	.41 (.61)*	.28 (.59)
2AFC recognition	.60 (.25)*	.28 (.35)*	.51 (.47)*	.29 (.46)*	.41 (.51)*	.49 (.42)*

Note. *For 2AFC only target-present trials are included. 2AFC = 2-alternative forced choice.*
*$p < .05$.

confidence–accuracy correlations for unstudied items that were highly similar to studied items, using yes–no recognition tests. Similarly, Sampaio and Brewer (2009) found negative confidence–accuracy correlations in a sentence recognition paradigm for a "deceptive" condition in which foil sentences were strongly consistent with the schema induced by the studied sentences. Rogers, Jacoby, and Sommers (2012) found negative confidence–accuracy correlations in the identification of spoken words presented amid noise when sensory and contextual information were incongruent with each other. Finally, Koriat (2012) obtained negative confidence–accuracy correlations for general knowledge questions that are often answered incorrectly.

The direction of the confidence–accuracy resolution (i.e., positive vs. negative) should influence whether allowing participants to choose which responses to volunteer or withhold increases or decreases their output-bound memory performance. That is, if resolution is positive, participants may improve their recognition performance when they are given the choice to volunteer or withhold a response, because they are generally right about when they are likely to be correct versus incorrect and thus can volunteer predominantly correct responses. On the other hand, if resolution is negative (e.g., with high-similarity lures), participants may worsen their recognition performance when they are given the choice to volunteer or withhold a response, because they are generally wrong about when they are likely to be correct versus incorrect and thus may end up volunteering predominantly incorrect responses. We test these predictions in Experiment 2.

EXPERIMENT 2

In Experiment 1 we found that face recognition performance was better for the 2AFC test format than for the yes–no test format. The purpose of Experiment 2 was to investigate whether this advantage extended to completely new lures, which would be analogous to a target-absent lineup in eyewitness identification. The target-absent condition is critical because it is the situation in which an innocent person can be wrongfully convicted. It is also important to include such a condition because lineup experiments derive their false alarm rates exclusively from target-absent lineups (see Note 2). Additionally, we were interested in applying Koriat and Goldsmith's (1994, 1996) technique of having people reflect a second time on their recognition responses and decide whether or not to volunteer them. Would the advantage of 2AFC to yes–no recognition remain under these conditions? Also, we added a manipulation of study duration, because exposure time has been shown to be an important variable in eyewitness memory performance (Shapiro & Penrod, 1986).

METHOD

Design and Participants

The experiment used a 2 (test format) × 2 (study duration) × 2 (lure similarity) × 2 (target presence or absence) design. Test format (yes–no vs. 2AFC) and

study duration (4 s vs. 8 s) were manipulated between subjects. Lure similarity to the target (high vs. low) and whether the target appeared on the test (target-present vs. target-absent) were manipulated within subjects. Dependent measures were recognition responses (yes–no or left–right), confidence ratings (50–100 for both yes–no and 2AFC), and report decisions (volunteer vs. withhold response). Participants were 72 undergraduate students (51 female, mean age 19 years, SD = 1) recruited through the Washington University participant pool who received either $10 or course credit for their participation.

Materials

Materials were 30 sets of face pictures, 23 of which were used in Experiment 1 and 7 of which were created anew using the same procedure described in Experiment 1. Again, each set consisted of three faces: an original, a high-similarity lure, and a low-similarity lure.

Procedure

The overall procedure was similar to that of Experiment 1. However, instead of making their responses using the keyboard as in Experiment 1, participants in Experiment 2 spoke all their responses out loud, and they were then entered into the computer via keyboard by a research assistant. Also, whereas in Experiment 1 a confidence scale of 0–100% was used for the yes–no test format and a scale of 50–100% was used for the 2AFC test format, in Experiment 2 the 50–100% scale was used for both test formats. Participants were instructed that 50% meant they were guessing and 100% meant they were absolutely sure they were correct.

The procedure again consisted of a study phase and a test phase. At the start of the study phase, participants were given more specific instructions than in Experiment 1:

> In the first part of the experiment you will view a series of faces on the screen. Imagine that the faces are pictures of known criminals who are wanted by the police. Your task is to study the faces carefully so that you can later identify them on a memory test.

Participants then viewed 20 of the 30 original faces at a rate of either 4 s per face (n = 36) or 8 s per face (n = 36), depending on the study duration condition to which each participant had been randomly assigned. The interstimulus interval was again 500 ms. Study order was randomized for each participant, and faces were counterbalanced so that all 30 original faces were used in the study list equally often across participants.

For the test phase, participants were randomly assigned to either the yes–no recognition test format (n = 36) or the 2AFC recognition test format (n = 36). Instructions at the beginning of both test formats described the procedure for each trial, including the following instructions regarding the report decision to be made for each trial (volunteer vs. withhold):

> Finally, for each face, you will be asked to indicate if you would like your response to count. Imagine that your response could be used to prosecute a suspect in court. In that case, you would only identify a suspect if you were quite certain that he was indeed the criminal that you saw. It is important that you identify criminals so that they can be put in jail. However, you would not want to put an innocent person in jail. Thus, you should only count your response if you are quite certain that you are correct.

YES–NO RECOGNITION TEST.

The procedure for each trial was similar to that in Experiment 1, with the addition of a report decision after the confidence judgment for each trial. After making their recognition decision (old vs. new) for a given trial and entering their confidence judgment for that trial, participants were given the prompt, "Do you want your response to count?" and they responded either "yes" or "no." There were 60 test trials total. The target-present condition comprised 40 trials: the 20 studied original faces and 20 lures, one corresponding to each studied original face (10 high-similarity lures and 10 low-similarity lures). The target-absent condition comprised 20 trials: the 10 unstudied original faces (lures) and 10 other "lures," one corresponding to each unstudied original face (5 high similarity to the unstudied original face and 5 low similarity to the unstudied original face).

Face sets were counterbalanced such that across participants each face set was equally represented in the two lure conditions (high similarity vs. low similarity) and in the studied versus unstudied conditions (target-present vs. target-absent). In the target-absent condition, the two lure faces bore either low or high similarity to one another but no similarity to any target face. Test trials occurred in an order that was randomized for each participant, with the constraints that two faces from the same set always occurred in

different halves of the test, and the original face occurred before its corresponding lure face exactly half of the time. The mean lag between original faces and their corresponding lures was 28.6 intervening trials (SD = 10.9, range 0–57).

2AFC RECOGNITION TEST.

As with the yes–no recognition test format, the procedure for each trial in the 2AFC recognition test format was the same as in Experiment 1, with the addition of the report decision after the confidence judgment for each trial. There were 30 trials in total. The target-present condition comprised 20 trials: 10 trials in which a studied original face was paired with its high-similarity lure and 10 trials in which a studied original face was paired with its low-similarity lure. The target-absent condition comprised 10 trials: 5 trials in which an unstudied original face (a lure) was paired with its high-similarity "lure" and 5 trials in which an unstudied original face was paired with its low-similarity "lure." Participants were informed that on some trials, neither of the faces would be ones that were studied but that they should pick one; however, in such cases they should choose to not have their response count (i.e., to withhold it) when they came to the report decision.

Face sets were counterbalanced such that across participants each face set was equally represented in the two lure conditions (high similarity vs. low similarity), in the studied versus unstudied conditions (target present vs. target absent), and in the position of the original face (left vs. right). In the target-absent condition, the two lure faces bore either low or high similarity to one another but no similarity to any target face. Test trials occurred in an order that was randomized for each participant, and in exactly half of the trials the original face (whether studied or unstudied) was on the left side.

Note that for both yes–no and 2AFC test formats, two faces were tested from all face sets, including the 20 sets for which the original face was studied and the 10 sets that were not studied at all. We use the terms "target-present" and "target-absent" to refer to each pair of faces on the tests, regardless of whether they were tested simultaneously (2AFC) or sequentially (yes–no).

RESULTS AND DISCUSSION

Collapsing across lure similarity, performance did not differ for the short versus the long presentation duration in the 2AFC test format, $t(34) = 1.28$, $p = .210$, $d = 0.43$, or in the yes–no test format, $t(34) = 0.39$, $p = .697$, $d = 0.13$. Thus, for all subsequent analyses we collapse across presentation duration.

Recognition Performance

We will first consider results from the target-present condition in one subsection and then results from the target-absent condition in a second subsection. Within each subsection we will separately consider what we call *full report performance* and *free report performance*. Full report performance includes all responses, regardless of participants' report decisions (volunteer vs. withhold), whereas free report performance includes only responses that participants decided to volunteer.

Target-Present

Performance data from the target-present condition are shown in Table 3.

FULL REPORT PERFORMANCE.

Full report performance data are shown in the top third of Table 3. Using AUC as the dependent variable, we essentially replicated the main finding of Experiment 1 in that the 2AFC test led to greater discriminability than the yes–no test. This point was confirmed in a two-way mixed ANOVA that revealed a main effect for test format, $F(1, 70) = 43.74$, $MSE = .019$, $p < .001$, $\hat{\omega}_p^2 = .373$. Performance was again better for items in the low-similarity condition versus the high-similarity condition, $F(1, 70) = 50.42$, $MSE = .014$, $p < .001$, $\hat{\omega}_p^2 = .224$. There was no reliable interaction between test format and lure similarity, $F(1, 70) = 0.01$.

FREE REPORT VOLUNTEER RATE.

Table 4 shows the means and standard deviations of the proportion of their responses that participants volunteered. In the target-present condition, participants in both test format conditions volunteered a greater proportion of their hits than their false alarms, suggesting that they had overall good insight into which of their responses were more likely to be correct. However, notice that this difference is quite diminished for high-similarity lures in the yes–no recognition test format. As anticipated in Experiment 1, the reason for this will become apparent when we consider confidence–accuracy relationships later in this *Results* section.

TABLE 3. Target-Present Condition: Mean (*SD*) Performance, Experiment 2

Report option × test format × lure similarity	Hit rate	False alarm rate	*d′*	AUC
Full report (all responses)				
Yes–no recognition				
Low similarity	.61 (.17)	.22 (.16)	1.16 (0.76)	.72 (.12)
High similarity	.61 (.21)	.49 (.21)	0.36 (0.61)	.58 (.11)
2AFC recognition				
Low similarity	.85 (.12)	.15 (.12)	2.15 (0.86)	.87 (.10)
High similarity	.71 (.17)	.29 (.17)	1.27 (1.10)	.74 (.16)
Free report (volunteered responses): input bound				
Yes–no recognition				
Low similarity	.37 (.18)	.06 (.08)	1.01 (0.63)	.65 (.10)
High similarity	.34 (.19)	.22 (.14)	0.34 (0.48)	.56 (.09)
2AFC recognition				
Low similarity	.61 (.20)	.04 (.06)	1.81 (0.67)	.78 (.11)
High similarity	.49 (.16)	.15 (.14)	1.08 (0.77)	.68 (.12)
Free report (volunteered responses): output bound				
Yes–no recognition				
Low similarity	.75 (.21)	.18 (.24)	1.39 (0.74)	.79 (.16)
High similarity	.73 (.31)	.54 (.29)	0.44 (0.65)	.60 (.18)
2AFC recognition				
Low similarity	.93 (.10)	.07 (.10)	2.42 (0.60)	.94 (.09)
High similarity	.78 (.18)	.22 (.18)	1.52 (1.00)	.80 (.17)

Note. *2AFC = 2-alternative forced choice; AUC = area under the curve.*

TABLE 4. Mean (*SD*) of Proportion of Responses Volunteered in Experiment 2

	Target present			Target absent	
Test format × lure similarity	All responses	Hits	False alarms	All responses	False alarms
Yes–no recognition					
Low similarity	.47 (.19)	.61 (.25)	.23 (.28)	.53 (.25)	.40 (.32)
High similarity	.44 (.16)	.53 (.27)	.50 (.30)	.45 (.21)	.32 (.31)
2AFC recognition					
Low similarity	.65 (.20)	.72 (.20)	.32 (.38)	.24 (.26)	
High similarity	.65 (.17)	.70 (.18)	.52 (.33)	.28 (.23)	

Note. *The "All responses" columns show the overall rate of volunteering calculated across all the types of responses possible for a given condition: for target-present yes–no: hits, misses, correct rejections, and false alarms; for target-present 2AFC: hits and false alarms; for target-absent yes–no: correct rejections and false alarms; for target-absent 2AFC: false alarms. False alarms were the only type of response possible in target-absent 2AFC, so the "All responses" and "False alarms" columns are combined in that case. 2AFC = 2-alternative forced choice.*

FREE REPORT PERFORMANCE.

In free report, there are two possible ways to evaluate memory performance: input bound ("quantity") and output bound ("accuracy").[4] Input-bound performance is typically calculated as the proportion of total trials on which a participant responded correctly and volunteered for that response to count. Output-bound performance is typically calculated as the proportion of a participant's volunteered responses that were in fact correct. For comparison to full report performance, we will focus our analyses here on input-bound free report performance. We will make use of the output-bound free report performance when we consider metacognition later in this *Results* section.

For the sake of clarity, we will provide an example of how we calculate performance measures for full report, free report input bound, and free report output bound. For yes–no recognition, there were 10 test trials that showed a studied (old) face in the high-similarity condition. Imagine that a participant correctly responded "yes" (old) on 7 of those 10 trials and incorrectly responded "no" (new) on 3 of those 10 trials. His *full report* hit rate for this condition would be 7/10 = .70. Imagine that he volunteered for 4 of his "yes" responses and 1 of his "no" responses to be counted. His free report input-bound hit rate for this condition would be 4/10 = .40. His free report output-bound hit rate for this condition would be 4/(4 + 1) = .80. The same basic approach to calculation applies in the case of false alarms and in the case of the 2AFC test format.

Input-bound free report data are shown in the middle third of Table 3. Once again, we find that 2AFC tests provide better discrimination than the yes–no test. Using AUC as the dependent variable, we conducted a three-way mixed ANOVA (report option × test format × lure similarity). Performance was again better for the 2AFC test format than for the yes–no test format, $F(1, 70) = 47.81$, $MSE = .030$, $p < .001$, $\hat{\omega}_p^2 = .394$. Performance was also better for items in the low-similarity condition than in the high-similarity condition, $F(1, 70) = 55.29$, $MSE = .019$, $p < .001$, $\hat{\omega}_p^2 = .227$. Finally, performance was better for full report versus free report responding by the input-bound scoring criterion, $F(1, 70) = 67.45$, $MSE = .004$, $p < .001$, $\hat{\omega}_p^2 = .091$. The only significant interaction was report option × lure similarity, $F(1, 70) = 7.29$, $MSE = .004$, $p = .009$, $\hat{\omega}_p^2 = .009$ (the effect of lure similarity was slightly smaller under free report but in the same direction). Overall, results showed the same pattern of performance as in Experiment 1, even when we allowed participants to volunteer or withhold their responses.

Target Absent

Performance data from the target-absent condition are shown in Table 5. The measures d' and AUC were calculated using the hit rates from the target-present condition (Table 3) and the false alarm rates from the target-absent condition. Note that *lure similarity* in the target-absent condition refers to the similarity of the two lure faces *to each other*; the two lures were not at all similar to any studied targets.

FULL REPORT PERFORMANCE.

Full report performance data for the target-absent condition are shown in the upper half of Table 5. For yes–no recognition, AUC performance was better for items in the low-similarity condition than in the high-similarity condition, although the difference was not statistically significant, $t(35) = 1.47$, $p = .151$, $d = 0.25$. We could not examine full report performance for 2AFC in the target-absent condition in this experiment because the procedure forced participants to choose one face or the other, even though neither had been studied, yielding a false alarm rate of 1. Experiment 3 will add an appropriate "neither" option that will allow us to analyze 2AFC full report target-absent performance.

FREE REPORT PERFORMANCE.

Free report performance data (input bound) for the target-absent condition are shown in the lower half of Table 5. Using AUC as the dependent variable, we conducted a two-way mixed ANOVA (test format × lure similarity). Performance was slightly better for the 2AFC test format than for the yes–no test format, $F(1, 70) = 4.80$, $MSE = .022$, $p = .032$, $\hat{\omega}_p^2 = .051$. Performance was again better for items in the low-similarity condition than in the high-similarity condition, $F(1, 70) = 8.94$, $MSE = .010$, $p = .004$, $\hat{\omega}_p^2 = .034$, and there was no significant interaction, $F(1, 70) = 2.07$, $MSE = .010$, $p = .155$, $\hat{\omega}_p^2 = .005$.

Interestingly, the target-absent free reported false alarm rate (collapsed across lure similarity) was reliably lower for the yes–no test format than for the 2AFC test format, $t(70) = 3.91$, $p < .001$, $d = 0.92$, an apparent reversal of the pattern in the target-present case. However, keep in mind that unlike the yes–no

TABLE 5. Target-Absent Condition: Mean (*SD*) Performance, Experiments 2 and 3

Report option × test format × lure similarity	Experiment 2			Experiment 3		
	False alarm rate	*d′*	AUC	False alarm rate	*d′*	AUC
Full report (all responses)						
Yes–no recognition						
Low similarity	.24 (.17)	1.10 (0.84)	.71 (.14)	.07 (.07)	1.68 (0.55)	.78 (.08)
High similarity	.33 (.25)	0.84 (1.16)	.66 (.21)	.11 (.11)	1.57 (0.74)	.76 (.11)
2AFC recognition						
Low similarity	1 (0)	n/a	n/a	.33 (.28)	1.02 (0.94)	.73 (.17)
High similarity	1 (0)	n/a	n/a	.38 (.28)	.66 (0.88)	.65 (.15)
Free report (volunteered responses): input bound						
Yes–no recognition						
Low similarity	.10 (.11)	0.90 (0.71)	.64 (.11)	.02 (.04)	1.41 (0.56)	.72 (.10)
High similarity	.12 (.15)	0.73 (0.95)	.61 (.15)	.04 (.07)	1.21 (0.55)	.68 (.09)
2AFC recognition						
Low similarity	.24 (.26)	1.00 (0.78)	.72 (.13)	.14 (.20)	0.91 (0.87)	.68 (.16)
High similarity	.28 (.23)	0.57 (0.62)	.64 (.11)	.10 (.14)	0.58 (0.67)	.63 (.11)

Note. d′ *and AUC were calculated using hits from the target-present condition, found in Tables 3 and 6. In Experiment 2, false alarm rate was 1 for full report 2AFC because participants were forced to choose between 2 unstudied faces. Lure similarity in the target-absent condition refers to the similarity of the 2 lure faces* to each other*; the 2 lures were not at all similar to any studied targets. 2AFC = 2-alternative forced choice; AUC = area under the curve.*

case, participants in the 2AFC target-absent case were required to guess a response even though they may have known that both were incorrect. It may well be that the process of forcing participants to respond makes them more likely to falsely believe later that their response is correct (e.g., Ackil & Zaragoza, 1998; Roediger, Jacoby, & McDermott, 1996). In Experiment 3, we sought to replicate this finding with an improved procedure for 2AFC target-absent trials (i.e., the inclusion of a "neither" option so that participants were not forced to respond).

METACOGNITION.

Table 2 shows the mean confidence–accuracy gamma correlations. For the 2AFC test format, target-absent trials were excluded because participants did not have the option of responding accurately on those trials. Other than that, correlations were calculated using data from all responses (i.e., disregarding volunteer–withhold report decisions). As in Experiment 1, the correlations were positive in every condition except for high-similarity lures in the yes–no test format, where the correlation was again negative (though not reaching statistical significance this time), $t(34) = 1.07$, $p = .291$. To examine the performance consequences of this pattern of correlations, we turn to output-bound free report measures, shown in the bottom third of Table 3. Assuming that participants' decisions to volunteer or withhold their responses were based at least in part on their confidence, their metacognitive resolution should be related to their output-bound free report memory performance. That is, if their metacognitive resolution is positive, then they should be more likely to withhold erroneous responses and thus increase their performance compared with the full report measures. And indeed, comparing the output-bound free report data in the bottom third of Table 3 (target-present condition) with the full report data in the top third of Table 3, we see that strategically withholding responses improved performance in every case *except* for the false alarm rate in yes–no recognition for the target-present high-similarity condition,

which *increased* from .49 to .54 while the comparable false alarm rate in 2AFC recognition decreased from .29 to .22. This interaction was statistically significant, $t(69) = 2.35$, $p = .022$, $d = 0.56$. It appears that a face highly similar to one studied previously gives rise to high-confidence false alarms when that face is viewed in isolation (yes–no) but not when it is viewed alongside the actually studied face (2AFC).

EXPERIMENT 3

In Experiment 2 we found that the advantage for 2AFC judgments over yes–no judgments extended to free report performance. However, we were unable to adequately assess whether the advantage extended to a target-absent condition, because participants were not given the option to respond correctly to target-absent trials in the 2AFC test format (i.e., they were forced to select one or the other of the faces even when neither had been studied). The main new feature introduced in Experiment 3 was the addition of a "neither" response option in the 2AFC test format, so that participants could respond correctly to target-absent trials. Note that although the addition of a "neither" option technically renders this test format not two-alternative forced-choice in the strict traditional sense, we nevertheless maintain the term *2AFC* for the sake of convenience and consistency with the first two experiments.

METHOD

Design and Participants

The experiment used a 2 (test format) × 2 (lure similarity) × 2 (target presence or absence) design. Test format (yes–no vs. 2AFC) was manipulated between subjects. Lure similarity to the target (high vs. low) and target presence on test (target-present vs. target-absent) were manipulated within subjects. The study duration manipulation used in Experiment 2 was dropped, because it had no effect in that experiment. Dependent measures were recognition responses (yes–no or left, right, or neither), confidence ratings (0–100 for both yes–no and 2AFC test formats), and report decisions (volunteer vs. withhold response). Participants were 50 undergraduate students (30 female, 1 unspecified, mean age 19 years, $SD = 1$) recruited through the Washington University participant pool who received either $10 or course credit for their participation.

Materials

Materials were the same 30 face sets used in Experiment 2.

Procedure

All participants in Experiment 3 were tested individually on computers programmed with Adobe Flash (Weinstein, 2012). Participants made their responses using the computer mouse to click on-screen buttons for recognition responses and report decisions and to click an on-screen slider for confidence values. Also, whereas in Experiment 2 a confidence scale of 50–100% was used for the both test formats, in Experiment 3 a 0–100% scale was used for both test formats. Participants were instructed that 0% meant they were purely guessing, and 100% meant they were absolutely sure they were correct.

The procedure again consisted of a study phase and a test phase. At the start of the study phase, participants were instructed that they would view a series of faces that they should imagine are pictures of criminals who have committed crimes around town. They were instructed that their task would be to study the faces carefully so that they could later help the police identify these criminals on a memory test. Participants viewed 20 of the 30 original faces at a rate of 4 s per face with a 500-ms interstimulus interval. Which particular faces were presented, and in what order, was randomized anew for each participant.

The composition of the tests, for both test formats, was the same as in Experiment 2, with the exception that randomization was used instead of counterbalancing. In the yes–no test the mean lag between original faces and their corresponding lures was 18.8 intervening trials ($SD = 13.6$, range 0–57). In the 2AFC test, a "neither" button was added on the screen between the "left" and "right" response buttons. Note that a "neither" response is analogous to rejecting the lineup in an eyewitness identification task. Participants were given the following instructions at the start of the test phase:

> Now you will complete a memory test for the faces that you studied. Imagine that the police have found a number of people who may have committed crimes. Some of these people are criminals, and some of them are innocent. The police need you to help identify which faces belong to the real criminals that you studied earlier.
>
> [Yes/No] You will see a face on the screen and you will decide if you studied that face ear-

lier. You will click a button to indicate that you did or did not study that face earlier.

[2AFC] You will see two faces on the screen and you will decide which one of them (if any) you studied earlier. You will click the button below the face you studied earlier, or a button indicating that neither were studied earlier.

Then, you will use a slider to rate your confidence in your answer, on a scale of 0% to 100%, where 0% means that you are purely guessing and 100% means that you are absolutely sure you are correct.

Finally, you will decide whether or not to officially report your answer to the police. Imagine that if you choose to report, your testimony will be used in a court of law. If you report that the person is a previously-studied criminal, he will likely be sent to jail for many years. If you report that the person is NOT a previously-studied criminal, he will likely go free. You will click a button to indicate whether you want to report your answer or not report your answer.

RESULTS AND DISCUSSION

Recognition Performance

Note that for the two-alternative test format, a "neither" response constituted either a miss in target-present trials or a correct rejection in target-absent trials, and thus such responses do not contribute to the calculation of hit rate, false alarm rate, d', or AUC.

Target Present

Performance data from the target-present condition are shown in Table 6.

TABLE 6. Target-Present Condition: Mean (*SD*) Performance, Experiment 3

Report option × test format × lure similarity	Hit rate	False alarm rate	d'	AUC
Full report (all responses)				
Yes–no recognition				
Low similarity	.60 (.16)	.14 (.12)	1.42 (0.56)	.75 (.09)
High similarity	.62 (.18)	.35 (.14)	0.76 (0.66)	.65 (.12)
2AFC recognition				
Low similarity	.70 (.16)	.06 (.07)	1.97 (0.62)	.83 (.09)
High similarity	.62 (.15)	.16 (.14)	1.39 (0.78)	.74 (.12)
Free report (volunteered responses): input bound				
Yes–no recognition				
Low similarity	.45 (.21)	.06 (.10)	1.25 (0.53)	.70 (.10)
High similarity	.40 (.20)	.20 (.17)	0.66 (0.63)	.61 (.10)
2AFC recognition				
Low similarity	.49 (.23)	.02 (.04)	1.54 (0.71)	.74 (.12)
High similarity	.35 (.22)	.05 (.07)	0.99 (0.65)	.65 (.11)
Free report (volunteered responses): output bound				
Yes–no recognition				
Low similarity	.73 (.25)	.12 (.23)	1.66 (0.66)	.82 (.13)
High similarity	.71 (.25)	.45 (.35)	0.77 (0.60)	.65 (.19)
2AFC recognition				
Low similarity	.83 (.21)	.03 (.05)	2.11 (0.69)	.90 (.12)
High similarity	.80 (.21)	.10 (.13)	1.34 (0.87)	.86 (.14)

Note. *2AFC = 2-alternative forced choice; AUC = area under the curve.*

FULL REPORT PERFORMANCE.

Full report performance data are shown in the top third of Table 6. Using AUC as the dependent measure, a two-way mixed ANOVA revealed that performance was again better for the 2AFC test format than for the yes–no test format, $F(1, 48) = 12.22$, $MSE = .015$, $p = .001$, $\hat{\omega}_p^2 = .183$. Performance was also again better when targets were accompanied by low-similarity lures than by high-similarity lures, $F(1, 48) = 25.92$, $MSE = .008$, $p < .001$, $\hat{\omega}_p^2 = .151$. There was no reliable interaction between test format and lure similarity, $F(1, 48) = 0.22$.

FREE REPORT VOLUNTEER RATE.

Table 7 shows the means and standard deviations of the proportion of their responses that participants volunteered. As in Experiment 2, in the target-present condition participants volunteered a greater proportion of their hits than their false alarms. Note that in the target-absent condition, the volunteer rate for false alarms was practically identical across test formats. That is, given that a participant had just made a false alarm, the participant was equally likely to volunteer that response whether she or he was in the yes–no or the 2AFC test format condition. However, just how many false alarms were made in the first place for the two test formats will be revealed when we consider free report performance.

FREE REPORT PERFORMANCE.

Input-bound free report performance was calculated as in Experiment 2. These data are shown in the middle third of Table 6. Using AUC as the dependent measure, we essentially replicated the results of Experiments 1 and 2. We conducted a three-way mixed ANOVA (report option × test format × lure similarity). Performance was again better for the 2AFC test format versus the yes–no test format, $F(1, 48) = 6.84$, $MSE = .029$, $p = .012$, $\hat{\omega}_p^2 = .104$. Performance was again better for items in the low-similarity condition than for the high-similarity condition, $F(1, 48) = 30.54$, $MSE = .013$, $p < .001$, $\hat{\omega}_p^2 = .159$. Performance was better for full report than for free report responding, $F(1, 48) = 63.66$, $MSE = .004$, $p < .001$, $\hat{\omega}_p^2 = .127$. The only significant interaction was report option × test format, $F(1, 48) = 7.41$, $MSE = .004$, $p = .009$, $\hat{\omega}_p^2 = .014$ (the effect of test format was slightly smaller under free report, but in the same direction).

Target Absent

Performance data from the target-absent condition are shown in Table 5.

FULL REPORT PERFORMANCE.

Full report performance data for the target-absent condition are shown in the upper half of Table 5. Using AUC as the dependent variable, a two-way mixed ANOVA revealed that performance was worse for the 2AFC test format than for the yes–no test format, $F(1, 48) = 6.95$, $MSE = .022$, $p = .011$, $\hat{\omega}_p^2 = .107$, because of the greater false alarm rate in the two-alternative test versus yes–no recognition when there was no target present, $t(48) = 5.30$, $p < .001$, $d = 1.48$. Note that

TABLE 7. Mean (*SD*) of Proportion of Responses Volunteered, Experiment 3

Test format × lure similarity	Target present			Target absent	
	All responses	Hits	False alarms	All responses	False alarms
Yes–no recognition					
Low similarity	.62 (.24)	.72 (.25)	.44 (.46)	.64 (.25)	.38 (.45)
High similarity	.56 (.24)	.66 (.27)	.54 (.42)	.52 (.27)	.37 (.40)
2AFC recognition					
Low similarity	.62 (.27)	.68 (.26)	.38 (.49)	.52 (.35)	.38 (.41)
High similarity	.47 (.29)	.56 (.30)	.42 (.44)	.46 (.30)	.37 (.42)

Note. *The "All responses" columns show the overall rate of volunteering calculated across all the types of responses possible for a given condition: for target-present yes–no: hits, misses, correct rejections, and false alarms; for target-present 2AFC: hits, misses, and false alarms; for target-absent yes–no: correct rejections and false alarms; for target-absent 2AFC: correct rejections and false alarms. 2AFC = 2-alternative forced choice.*

this outcome occurred even though participants had the option to say "neither" (i.e., to reject the pair of lures). This striking reversal between the superiority of yes–no and the 2AFC test is clearly illustrated in Figure 2, in the false alarm rates for target present versus target absent. Performance was only marginally better for items in the low-similarity condition than in the high-similarity condition, $F(1, 48) = 3.81$, $MSE = .015$, $p = .057$, $\hat{\omega}_p^2 = .022$. There was no reliable interaction between test format and lure similarity, $F(1, 48) = 1.40$, $MSE = .015$, $p = .242$, $\hat{\omega}_p^2 = .003$.

FREE REPORT PERFORMANCE.

Free report performance data (input-bound) for the target-absent condition are shown in the lower half of Table 5. Using AUC as the dependent measure, a three-way mixed ANOVA revealed that performance was again worse for the 2AFC test format than for the yes–no test format, $F(1, 48) = 5.32$, $MSE = .036$, $p = .025$, $\hat{\omega}_p^2 = .079$, because of the high false alarm rates in 2AFC. Performance was again better for items in the low-similarity condition than in the high-similarity condition, $F(1, 48) = 4.54$, $MSE = .023$, $p = .038$, $\hat{\omega}_p^2 = .027$. Performance was better for full report than for free report scoring, $F(1, 48) = 29.29$, $MSE = .005$, $p < .001$, $\hat{\omega}_p^2 = .065$. There were no significant three-way or two-way interactions.

The key finding here is that although the 2AFC test format yielded better performance than the yes–no test format when a studied target was present, it yielded worse performance when both faces were new, and that allowing participants to strategically withhold responses did not change this pattern. Note specifically that the target-absent false alarm rates are higher for 2AFC than yes–no, $t(48) = 3.16$, $p = .003$, $d = 0.88$, replicating the result from Experiment 2 even with the option provided for participants to respond "neither" in Experiment 3 in the full report condition.

METACOGNITION.

Table 2 shows the mean confidence–accuracy gamma correlations. For the 2AFC test format, target-absent trials were excluded because there were too few observations per participant per cell (5) to calculate stable estimates of gamma separately from the target-present trials (Spellman, Bloomfield, & Bjork, 2008). Other than that, correlations were calculated using data from all responses (i.e., disregarding volunteer or withhold report decisions). As in Experiments 1 and 2, the correlation was again negative for high-similarity lures in the yes–no test format, although not statistically significantly so, $t(24) = 0.54$, $p = .591$. But the consistency of this null or negative correlation across all three experiments is nevertheless compelling in comparison to all other conditions. Furthermore, combining data from all three experiments for this particular cell yielded a mean correlation of –.11

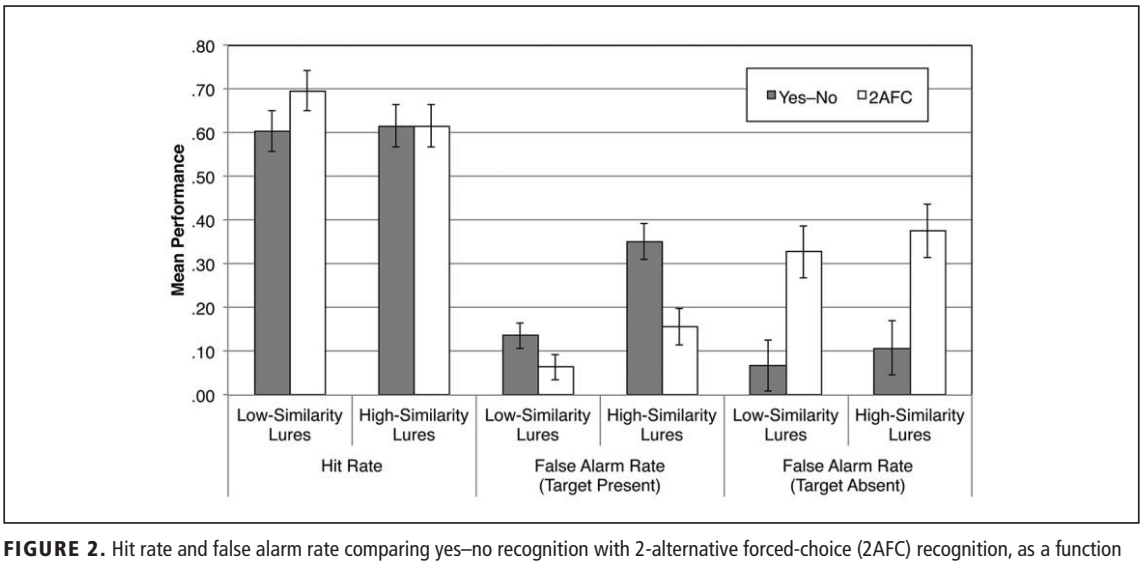

FIGURE 2. Hit rate and false alarm rate comparing yes–no recognition with 2-alternative forced-choice (2AFC) recognition, as a function of lure similarity and target presence, Experiment 3. Data are full report (all responses). Error bars represent the pooled *SE* for the between-subject comparison of yes–no and 2AFC

(SD = .52), $t(79)$ = 1.94, p = .057, and also revealed that the correlation was negative for a majority of the participants for whom it could be calculated in this cell (45 out of 80). We can again evaluate the consequences of that negative resolution by comparing the output-bound free report data in the bottom third of Table 6 (target-present condition) with the full report data in the top third of Table 6. We again see that strategically withholding responses improved or did not change performance in every case except for the false alarm rate in yes–no recognition for the target-present high-similarity condition, which increased from .35 to .45 while the comparable false alarm rate in 2AFC recognition decreased from .16 to .10. This interaction was statistically significant, $t(46)$ = 2.52, p = .015, d = 0.73.

GENERAL DISCUSSION

Four main findings emerged from our experiments. First, when a studied target was present among alternatives in the test, face recognition was superior in the simultaneous (2AFC) test format compared with the yes–no (sequential) test format in all three experiments. This outcome occurred under both full report and free report responding. Second, this pattern reversed when there was no target present in Experiment 3, despite the fact that participants could say that neither face was present (they could "reject" the choice). In addition, the button to respond "neither" was located on the computer screen between the left and right buttons used to make old responses, so the option was salient. Third, in target-present conditions, participants' confidence–accuracy resolution was positive in all circumstances except for high-similarity lures judged in isolation (yes–no recognition), which led to an increased output-bound false alarm rate in that condition. Fourth, permitting participants to withhold responses improved accuracy only under certain conditions. We discuss these findings in turn.

First, our finding that 2AFC tests yielded superior target-present discriminability to yes–no tests both conflicts with the prediction made by signal detection theory (Experiment 1) and concurs with the recent lineup experiments cited earlier. All three of our experiments confirmed this pattern. The top half of Figure 3 shows the mean ROC curves from the target-present condition in Experiment 3. The curves illustrate the substantial superiority of the simultaneous procedure in these experiments, whether lures were of low or high similarity to the target (left vs. right panel). This result adds to the debate questioning the wisdom of advising police departments to switch from simultaneous to sequential lineup procedures (Gronlund, Wixted, & Mickes, 2014).

The difference in performance between a simultaneous and sequential procedure may result from a difference in the basis for decisions used by participants in the two test formats. Wixted and Mickes (2014) proposed a diagnostic feature detection model of eyewitness identification to explain this result. Briefly, in lineup situations, eyewitnesses view multiple people who share some features in common (all fit the general verbal description given by the eyewitness) but others who are distinctive to the perpetrator (in a target-present lineup). In a simultaneous lineup, the eyewitness can view all the candidates and thus immediately discern that common features among them that can be discounted (e.g., if all have brown hair and brown eyes, that feature can be discounted). Thus they are more able to focus on distinctive features that might lead to accurate recognition. On the other hand, in the sequential lineup, with candidates presented one at a time, all features are possible candidates for distinctive features and, at least until near the end of the sequence, eyewitnesses may be cautious and fail to identify a suspect. This process of being cautious about the distinctive features in sequential presentations may account for the conservative criterion shift induced by sequential lineups. Because homing in on the correct distinctive features relevant to recognition is easier in the simultaneous lineup, greater accuracy of simultaneous lineups when targets are present is to be expected.

Yet what could explain our second main finding, that a simultaneous procedure was worse than a sequential procedure in the target-absent case in Experiment 3? This reversal is illustrated in the ROC curves shown in the bottom half of Figure 3. As just discussed, comparing the features of two faces seen at once is an effective way of determining which of the two was previously studied, if one of them was in fact previously studied. When both faces are new, it may be that the process of comparing the two faces gives rise to an inappropriate feeling of familiarity.

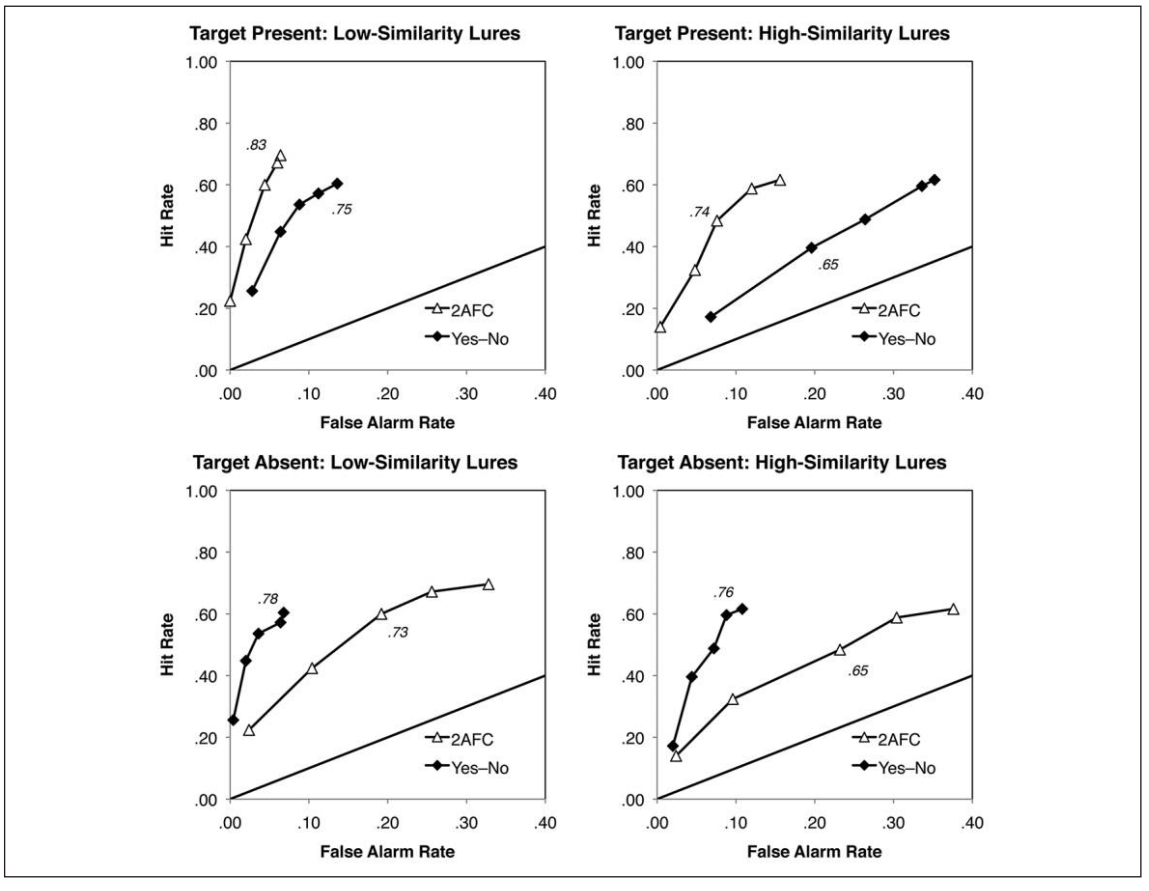

FIGURE 3. Receiver operating characteristics averaged across participants, comparing yes–no recognition with 2-alternative forced-choice (2AFC) recognition, as a function of lure similarity to target (low vs. high) and target presence (present vs. absent), Experiment 3. Data are full report (all responses). Italicized numbers are the mean area under the curve

Keep in mind that in our target-absent case (unlike most lineup research) the two faces that were seen were similar to each other but not similar to faces that had been studied. When participants look at the two similar faces in the simultaneous lineup, the perception of similarity may drive a false sense of familiarity (see Tulving, 1981, and Chandler, 1994, for related observations). That is, in the process of deciding which of two faces is more familiar, participants may fail to realize that neither is truly familiar. This process does not occur in the yes–no test, because the two lures are separated by faces from other lineups, so their similarity to each other is not as obvious. Note that the separation of related faces also makes our yes–no procedure different from most lineup experiments.

In sum, for target-present pairs, the obviousness of the difference between the target face and the lures increases discrimination of the target in simultaneous presentation of faces, in line with claims by Wixted and Mickes (2014). For target-absent pairs, the increased similarity (familiarity) induced by our using two lures that are similar to one another results in the judgment that one member of the pair was earlier presented, increasing false alarms. Prior research comparing simultaneous and sequential lineups has shown that a lineup including a person similar to the perpetrator is sufficient to eliminate the advantage of the simultaneous procedure (Mickes et al., 2012). We hypothesize that similarity between alternatives in a target-absent lineup can reverse the advantage of the simultaneous lineup and show superior discriminability for the sequential lineup.

Of course, several differences exist between the procedures used in lineup experiments and those

used in the current experiments. Encoding was intentional in our experiments and is often incidental in lineup experiments. We presented pairs of faces in our simultaneous test, whereas 6 or more faces are used in lineup experiments (and in real-world lineups). In addition, we presented many target faces for study (40 in Experiment 1, 20 in Experiments 2 and 3), whereas lineup experiments typically have only one target; and we tested many faces (80 in Experiment 1; 60 in Experiments 2 and 3), whereas in lineup experiments there is a test for but a single item in the target-present case (and 5 lures). A less obvious difference is that in Experiments 2 and 3, two thirds of the test pairs contained a target, so only one third of the pairs involved a target-absent condition. The relative proportion of target-present lineups is often not specified in experiments comparing simultaneous and sequential lineups.

Perhaps our using a preponderance of target-present tests biased participants toward mistakenly choosing a member of target-absent pairs. What is the case for true lineups? It is inherently impossible to answer this question with certainty, but eyewitnesses are likely to believe that the probability of the real perpetrator being present is quite high, in part because of the assumption that police will bother arranging a lineup only if they have strong suspicions that they have caught a guilty suspect. Brewer, Keast, and Rishworth (2002, p. 47) reported the beliefs of two seasoned and well-educated South Australian police detectives that only 10% of real-world lineups do not contain the actual perpetrator (N. Brewer, personal communication, July 16, 2014). Memon, Gabbert, and Hope (2004, p. 107) conducted four eyewitness lineup identification experiments with a total of 636 participants and found that 90% of them reported having assumed that the real perpetrator was indeed in the lineup to which they had just responded, despite the fact that they were given clear instructions that the perpetrator might not be present (and more than 90% of them recalled those cautionary instructions).

Our high-similarity faces were almost certainly more similar to each other than would be faces used to produce a "fair" or "unbiased" lineup (one in which no single alternative conspicuously matches the target more than the others do). But how similar should faces be to make a lineup truly fair? The guidelines for producing fair lineups include the instruction that all people in the lineup should match the general verbal description that the eyewitness has provided about the perpetrator. We doubt that real lineups involve similarity of lures as great as for our target-absent test pairs, but including such similar pairs may produce a fairer lineup by siphoning off false responses, as long as the two similar faces were known innocents. On the other hand, it is possible that making a lineup that consists of faces that are *too* similar to each other may mislead eyewitnesses into making more false alarms, particularly in a simultaneous procedure. Although we did not find such an interaction between lure similarity and test format in the target-absent conditions in our experiments, the possibility is worthy of additional research.

Finally, it is worth noting that the faces in our target-absent condition did not by design bear any particular resemblance to any of the studied faces, in contrast even to laboratory lineup experiments in which the target-absent lineups at least match a general verbal description of a perpetrator.

In short, the differences between our procedure and the standard lineup procedures may account for why our results differ from those of Mickes et al. (2012) and others, and these differences are sufficient to call into question the relevance of the current results to conclusions about the choice between lineup procedures. However, we believe our results at least should encourage additional research examining the relevance of similarity between lures as a factor in lineups, particularly simultaneous lineups. Although the importance of the similarity between the target and lures is obvious, a potential role played by similarity between lures is less obvious. Wells and Seelau (1995) reviewed evidence that certain lineup practices can particularly encourage relative judgments leading to false identification, and they made the recommendation that the suspect provided by police (who may or may not be the real perpetrator) "should not stand out in the lineup or photospread as being different from the distractors on the basis of the eyewitness's previous description of the culprit or other factors that would draw extra attention to the suspect" (p. 779). Thus, the role of variance between a suspect and the known innocent fillers in a lineup has been examined in prior research. We argue that the role of variance between even the known innocent alter-

natives in a lineup, given that they all equally fit the verbal description, is worthy of additional research.

Our third main finding is that when participants had to choose between target faces and other faces that had many features in common with the target face, they made many false alarms, and in the yes–no test format their confidence ratings were either uncorrelated or negatively correlated with accuracy of their judgments. This negative resolution has been found in other contexts, such as in the cases of foil sentences that were deceptively similar to originally studied sentences (Sampaio & Brewer, 2009), tricky general knowledge questions that many people miss (Koriat, 2012), and words that were semantically similar to those studied in a categorized list (DeSoto & Roediger, 2014; Roediger & DeSoto, 2014). To our knowledge, however, this is the first report of such negative resolution in face recognition. This outcome points to likely metacognitive problems in lineup experiments when a lure item is similar to the target, a problem that has long been appreciated in the eyewitness lineup identification literature (Buckhout, 1974).

Finally, our fourth main finding was that a variant of the procedure described by Koriat and Goldsmith (1996; Goldsmith & Koriat, 2008), in which participants are given the ability to volunteer or withhold a response after it has been made, aided performance only in certain conditions. Allowing such a report decision is most likely to be useful in situations where a premium is placed on output-bound performance (accuracy) over input-bound performance (quantity)—that is, when "telling nothing but the truth" is more important than "telling the whole truth." However, as illustrated in Experiments 2 and 3 by the consequences of the negative confidence–accuracy resolution for target-present lures in yes–no recognition, allowing participants to strategically volunteer or withhold their responses will improve output-bound performance only to the extent that their metacognition is effective.

Conclusion

Our experiments, using basic laboratory face recognition methods, show that simultaneous presentation of alternatives generally leads to better discriminability than does sequential presentation, when a target is among the alternatives. However, we also showed that when there is no target among the alternatives (target-absent) and the two lures resemble each other to some degree, a reversal occurs and the sequential procedure yields greater discriminability than the simultaneous procedure. Although our experimental procedures differ from those used in most lineup research (as well as real lineups), we believe that this outcome should spur additional research into the role of similarity of lures in lineup research.

NOTES

Andrea D. Hughes is now at Department of Psychology, University of the Fraser Valley.

This research was supported by a James S. McDonnell Foundation 21st Century Science Initiative in Bridging Brain, Mind, and Behavior Collaborative Award to Larry L. Jacoby and Henry L. Roediger III. We thank Chad Rogers for data management, Tammy Duguid for stimulus development, Carole Jacoby and Nancy Byars for data collection, and John Wixted and Ian Dobbins for invaluable advice on data analysis.

Address correspondence about this article to Jason R. Finley, Department of Psychology, Washington University, 1 Brookings Dr., St. Louis, MO 63130 (e-mail: jrfinley@wustl.edu).

1. In Experiments 1 and 3 an additional between-subject condition was run that consisted of 2AFC in which each target face on the test was paired with lures that were of low or high similarity to other target faces rather than low or high similarity to that target face itself. In all three experiments, an additional group of older adults was also run. Results from the alternative 2AFC condition and results from the older adults neither informed nor conflicted with the results from the main two test conditions and results from the undergraduate participants, and thus we do not report them here.

2. Note that whereas we use the full AUC, a partial version of AUC (pAUC) is used in analyzing performance from lineup procedures because there the maximum false alarm rate is limited to $1/m$, where m is the number of people in the lineup. Real-world lineups consist of one suspect and $m - 1$ known innocents. If an eyewitness identifies a known innocent, that response is disregarded by police. Lineup experiments parallel this situation by considering a response to be a false alarm only if it is an identification of a designated innocent suspect in a target-absent condition. When there is a target-absent condition with no designated innocent suspect, the false alarm rate is divided by m.

3. Note that the within-subject gamma correlation is a measure of metacognitive resolution: the general tendency to give higher confidence ratings to more accurate responses (cf. Smith, Kassin, & Ellsworth, 1989). Brewer and Wells (2006) argue that calibration is more informative for practical purposes in an eyewitness identification context, where only one judgment is made per eyewitness, on the grounds that a correlation could be low even when calibration is high

(Juslin, Olsson, & Winman, 1996). We will not be analyzing calibration because we think it is less relevant than resolution to understanding what processes give rise to differing levels of performance in our tasks. Furthermore, calculating a stable estimate of the calibration index would require hundreds of observations per participant, which we do not have, or else collapsing across participants, which ignores an important source of variance.

4. In full report, input-bound and output-bound performance are identical.

REFERENCES

Ackil, J. K., & Zaragoza, M. S. (1998). Memorial consequences of forced confabulation: Age differences in susceptibility to false memories. *Developmental Psychology, 34,* 1358–1372. doi:10.1037/0012-1649.34.6.1358

Andersen, S. M., Carlson, C. A., Carlson, M. A., & Gronlund, S. D. (2014). Individual differences predict eyewitness identification performance. *Personality and Individual Differences, 60,* 36–40. doi:10.1016/j.paid.2013.12.011

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,* 44–56. doi:10.1037/1076-898X.8.1.44

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

Buckhout, R. (1974). Eyewitness testimony. *Scientific American, 231*(6), 23–31.

Carlson, C. A., & Carlson, M. A. (2014). An evaluation of lineup presentation, weapon presence, and a distinctive feature using ROC analysis. *Journal of Applied Research in Memory and Cognition, 3*(2), 45–53. doi:10.1016/j.jarmac.2014.03.004

Chandler, C. C. (1994). Studying related pictures can reduce accuracy, but increase confidence, in a modified recognition test. *Memory & Cognition, 22,* 273–280.

Deffenbacher, K. A., Leu, J. R., & Brown, E. L. (1981). Memory for faces: Testing method, encoding strategy, and confidence. *American Journal of Psychology, 94,* 13–26. doi:10.2307/1422340

DeSoto, K. A., & Roediger, H. L. (2014). Positive and negative correlations between confidence and accuracy for the same events in recognition of categorized lists. *Psychological Science, 25,* 781–788. doi:10.1177/0956797613516149

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,* 345–357. doi:10.1037/a0034596

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

Goldsmith, M., & Koriat, A. (2008). The strategic regulation of memory accuracy and informativeness. In A. Benjamin & B. Ross (Eds.), *Psychology of learning and motivation, Vol. 48: Memory use as skilled cognition* (pp. 1–60). San Diego, CA: Elsevier.

Green, D. M., & Moses, F. L. (1966). On the equivalence of two recognition measures of short-term memory. *Psychological Bulletin, 66,* 228–234. doi:10.1037/h0023645

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics.* New York, NY: Wiley.

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

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

Interquest Inc. (1998). FACES: The Ultimate Composite Picture (3.0) [software application]. Houston, TX: IQ Biometrix.

Jang, Y., Wixted, J. T., & Huber, D. E. (2009). Testing signal-detection models of yes/no and two-alternative forced-choice recognition memory. *Journal of Experimental Psychology: General, 138,* 291–306. doi:10.1037/a0015525

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,* 1304–1316. doi:10.1037/0278-7393.22.5.1304

Koriat, A. (2012). The self-consistency model of subjective confidence. *Psychological Review, 119,* 80–113. doi:10.1037/a0025648

Koriat, A., & Goldsmith, M. (1994). Memory in naturalistic and laboratory contexts: Distinguishing the accuracy-oriented and quantity-oriented approaches to memory assessment. *Journal of Experimental Psychology: General, 123,* 297–315.

Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review, 103,* 490–517.

Kroll, N. E., Yonelinas, A. P., Dobbins, I. G., & Frederick, C. M. (2002). Separating sensitivity from response bias: Implications of comparisons of yes–no and forced-choice tests for models and measures of recognition memory. *Journal of Experimental Psychology: General, 131,* 241–254. doi:10.1037/0096-3445.131.2.241

Lindsay, R. C. L., & Wells, G. L. (1985). Improving eyewitness identifications from lineups: Simultaneous versus sequential lineup presentation. *Journal of Applied Psychology, 70,* 556–564. doi:10.1037/0021-9010.70.3.556

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). New York, NY: Cambridge University Press.

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (2nd ed.). Mahwah, NJ: Erlbaum.

McNicol, D. (2004). *A primer of signal detection theory*. Mahwah, NJ: Erlbaum.

Memon, A., Gabbert, F., & Hope, L. (2004). The ageing eyewitness. In J. Adler (Ed.), *Forensic psychology: Debates, concepts and practice* (pp. 96–112). Devon, UK: Willan.

Mickes, L., Flowe, H. D., & Wixted, J. T. (2012). Receiver operating characteristic analysis of eyewitness memory: Comparing the diagnostic accuracy of simultaneous and sequential lineups. *Journal of Experimental Psychology: Applied, 18,* 361–376. doi:10.1037/a0030609

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

Pollack, I., & Hsieh, R. (1969). Sampling variability of the area under the ROC-curve and of d'_e. *Psychological Bulletin, 71,* 161–173. doi:10.1037/h0026862

Psychology Software Tools. (2002). E-Prime (1.0) [software application]. Sharpsburg, PA: Author.

Roediger, H. L., & DeSoto, K. A. (2014). Confidence and memory: Assessing positive and negative correlations. *Memory, 22,* 76–91. doi:10.1080/09658211.2013.795974

Roediger, H. L., Jacoby, D., & McDermott, K. B. (1996). Misinformation effects in recall: Creating false memories through repeated retrieval. *Journal of Memory and Language, 35,* 300–318. doi:10.1006/jmla.1996.0017

Rogers, C. S., Jacoby, L. L., & Sommers, M. (2012). Frequent false hearing by older adults: The role of age differences in metacognition. *Psychology and Aging, 27,* 33–45. doi:10.1037/a0026231

Sampaio, C., & Brewer, W. F. (2009). The role of unconscious memory errors in judgments of confidence for sentence recognition. *Memory & Cognition, 37,* 158–163. doi:10.3758/MC.37.2.158

Shapiro, P., & Penrod, S. (1986). A meta-analysis of facial identification studies. *Psychological Bulletin, 100,* 139–156. doi:10.1037/0033-2909.100.2.139

Smith, V. L., Kassin, S. M., & Ellsworth, P. C. (1989). Eyewitness accuracy and confidence: Within- versus between-subjects correlations. *Journal of Applied Psychology, 74,* 356–359. doi:10.1037/0021-9010.74.2.356

Spellman, B. A., Bloomfield, A., & Bjork, R. A. (2008). Measuring memory and metamemory: Theoretical and statistical problems with assessing learning (in general and using gamma (in particular) to do so. In J. Dunlosky & R. A. Bjork (Eds.), *A handbook of metamemory and memory* (pp. 95–114). New York, NY: Psychology Press.

Steblay, N. K., Dysart, J. E., & Wells, G. L. (2011). Seventy-two tests of the sequential lineup superiority effect: A meta-analysis and policy discussion. *Psychology, Public Policy, and Law, 17,* 99–139. doi:10.1037/a0021650

Tulving, E. (1981). Similarity relations in recognition. *Journal of Verbal Learning and Verbal Behavior, 20,* 479–496. doi:10.1016/S0022-5371(81)90129-8

Weinstein, Y. (2012). *Flash programming for the social & behavioral sciences: A simple guide to sophisticated online surveys and experiments*. Thousand Oaks, CA: Sage.

Wells, G. L. (1984). The psychology of lineup identifications. *Journal of Applied Social Psychology, 14,* 89–103. doi:10.1111/j.1559-1816.1984.tb02223.x

Wells, G. L. (2014). Eyewitness identification: Probative value, criterion shifts, and policy regarding the sequential lineup. *Current Directions in Psychological Science, 23,* 11–16. doi:10.1177/0963721413504781

Wells, G. L., & Seelau, E. P. (1995). Eyewitness identification: Psychological research and legal policy on lineups. *Psychology, Public Policy, and Law, 1,* 765–791. doi:10.1037/1076-8971.1.4.765

Wells, G. L., Small, M., Penrod, S., Malpass, R. S., Fulero, S. M., & Brimacombe, C. A. E. (1998). Eyewitness identification procedures: Recommendations for lineups and photospreads. *Law and Human Behavior, 22,* 603–647. doi:10.1023/A:1025750605807

Wixted, J. T., & Mickes, L. (2014). A signal-detection–based diagnostic feature-detection model of eyewitness identification. *Psychological Review, 121,* 262–276. doi:10.1037/a0035940