



Identifying the guilty word: Simultaneous versus sequential lineups for DRM word lists

Jason R. Finley¹ · John T. Wixted² · Henry L. Roediger III³

Published online: 28 March 2020
© The Psychonomic Society, Inc. 2020

Abstract

Recent research in the eyewitness identification literature has investigated whether simultaneous or sequential lineups yield better discriminability. In standard eyewitness identification experiments, subjects view a mock-crime video and then are tested only once, requiring large samples for adequate power. However, there is no reason why theories of simultaneous versus sequential lineup performance cannot be tested using more traditional recognition memory tasks. In two experiments, subjects studied DRM (Deese-Roediger-McDermott) word lists (e.g., *bed*, *rest*, *tired*, ...) and were tested using “lineups” in which six words were presented either simultaneously or sequentially. A studied word (e.g., *tired*) served as the guilty suspect in target-present lineups, unstudied related words (e.g., *nap*) served as fillers in target-present and target-absent lineups, and critical lures (e.g., *sleep*) were included in some target-present and target-absent lineups as well, to serve as attractive alternatives to the target word (or suspect). ROC analyses showed that the simultaneous test format generally yielded superior discriminability performance compared to the sequential test format, whether or not the critical lure was present in the lineup.

Keywords Word recognition · Eyewitness memory · ROC · DRM · Methodology

Introduction

Experimental psychologists have long been concerned about the reliability of eyewitness memory (e.g., Clifford & Bull, 1978; Munsterberg, 1908), especially with regard to the identification of a suspect in a lineup. Few recognition memory tasks have consequences that are as important as a police lineup. Guilty suspects have been correctly convicted based on true eyewitness identifications, but innocent suspects have also been wrongly convicted based on false eyewitness identifications (Buckhout, 1974; Garrett, 2011; see also The Innocence Project <http://www.innocenceproject.org> and the National Registry of Exonerations, Gross & Shaffer, 2012).

Lineups are used for eyewitness identification purposes when the police establish a suspect during the course of a

criminal investigation. Importantly, the suspect may or may not actually be the one who committed the crime. An eyewitness who saw the crime is shown a lineup, which typically consists of a set of photographs with one photograph of the suspect and five or more photographs of physically similar “fillers” who are “known innocents” and not suspected of the crime (Police Executive Research Forum, 2013, Fig. 2). The eyewitness can either choose one of the lineup members or decline to choose anyone (i.e., they can “reject the lineup”).

This real-world procedure is both similar to and different from basic laboratory experiments on recognition memory, so it is worth clarifying a few points. In real-world lineups, police do not know when a suspect is guilty. In experiments, the researchers do indeed know when a suspect is guilty (i.e., was previously shown to subjects). If the suspect is guilty, then the lineup is a *target-present lineup*, and choosing the suspect is a *hit*, while rejecting the lineup is a *miss*. If the suspect is innocent, then the lineup is a *target-absent lineup*, and choosing the suspect is a *false alarm*, while rejecting the lineup is a *correct rejection*. If the eyewitness chooses a filler (i.e., any candidate beside the suspect), that is a different kind of false alarm, called a filler identification or “foil identification” (Clark & Godfrey, 2009), which carries much less consequence than a false alarm to an innocent suspect. Because fillers are known innocents, a filler identification will not

✉ Jason R. Finley
jfinley@fontbonne.edu

¹ Department of Biological and Behavioral Sciences, Fontbonne University, 6800 Wydown Blvd, St. Louis, MO 63105, USA

² Department of Psychology, University of California, San Diego, CA, USA

³ Department of Psychological and Brain Sciences, Washington University in St. Louis, St. Louis, MO, USA

result in investigation or prosecution of the person pictured. Although both filler identifications and lineup rejections (which could be misses or correct rejections) provide some evidence in favor of the suspect's innocence (Wells, Smalarz, & Smith, 2015; Wells, Smith, & Smalarz, 2015), they are less consequential than the responses that matter the most: when the eyewitness chooses the suspect, known as "suspect identifications" (IDs). In real-world settings, when an eyewitness makes a suspect ID, that suspect is likely to be further investigated and possibly charged with and convicted of the crime. Suspect ID responses are either hits or false alarms, depending on whether the suspect is actually guilty. The ideal lineup procedure would maximize correct suspect IDs (hits) and minimize incorrect suspect IDs (false alarms). Thus, much research has focused on these two measures, focusing on analyses that exclude filler identifications and lineup rejections (e.g., diagnosticity ratios, receiver operating characteristics (ROCs)).

Researchers have investigated numerous variables to determine their effect on eyewitness memory performance (e.g., Wells et al., 1998). One variable of particular interest is lineup presentation format: *simultaneous* versus *sequential*. In a simultaneous lineup, all six photos are presented at the same time and the eyewitness either chooses one of the candidates or rejects the lineup. In a sequential lineup, each photo is presented individually, one at a time, and the eyewitness makes a yes/no decision about each candidate. A "stopping rule" based on the first identification may be used, such that the sequential lineup ends as soon as the eyewitness chooses a candidate. If such a stopping rule is not used, the sequential lineup continues through all photos and the eyewitness may choose multiple candidates. Either way, the eyewitness can say no to all candidates, thus rejecting the lineup. Sequential lineups can also make use of a technique called backloading, in which subjects do not know exactly how many photos they will see (Horry, Palmer, & Brewer, 2012). Note that the sequential lineup procedure resembles an old/new (yes/no) recognition task, while the simultaneous lineup procedure resembles an n-alternative forced choice recognition task. Finley, Roediger, Wahlheim, Hughes, and Jacoby (Finley, Roediger, Hughes, Wahlheim, & Jacoby, 2015) provided further discussion of lineup procedures versus traditional laboratory recognition procedures.

A landmark study by Lindsay and Wells (1985) compared simultaneous versus sequential photo lineups and found that although there was a small and non-significant difference in the hit rate (simultaneous = .58, sequential = .50), the false-alarm rate was substantially lower for the sequential procedure (simultaneous = .43, sequential = .17). Intuitively, the larger proportional decrease in the false-alarm rate compared to the hit rate would seem to argue in favor of the sequential procedure. This apparent "sequential superiority effect" is quantified by the *diagnosticity ratio*, which is equal to the hit rate

divided by the false-alarm rate. Thus, in the Lindsay and Wells (1985) study, the diagnosticity ratio for the sequential lineup was $.50/.17 = 2.94$ and for the simultaneous lineup it was $.58/.43 = 1.35$.¹ The same pattern has often been observed in later investigations (Stebly, Dysart, & Wells, 2011).

According to one theory (Wells, 1984, 2001), simultaneous lineups encourage eyewitnesses to use a *relative* judgment process to select the candidate who looks most like the perpetrator they remember seeing. Some researchers have argued that the use of a relative judgment process tends to yield false identifications when the true perpetrator is not in the lineup, particularly when the lineup is unfair in that the innocent suspect resembles the perpetrator more than the fillers do (Lindsay et al., 1991). In contrast, sequential lineups theoretically encourage an *absolute* judgment process in which eyewitnesses individually compare each candidate to their memory of the perpetrator, and this process is less likely to lead to false identification (Wells, 1984, 2001). Based on empirical data indicating that the diagnosticity ratio is higher for sequential lineups, and on the theory of absolute versus relative judgments, as many as 30% of police departments in the USA adopted the sequential procedure (Police Executive Research Forum, 2013).

Recently, researchers have pointed out that the diagnosticity ratio would increase even if the only effect of sequential lineups was to induce more conservative responding (Gronlund, Mickes et al., 2014; Meissner, Tredoux, Parker, & MacLin, 2005; Mickes, Flowe, & Wixted, 2012; Wixted & Mickes, 2012). Indeed, as first noted by Wells (1984), the absolute/relative distinction speaks to *response bias* – the tendency to make an ID from the lineup at all – with the relative judgment strategy yielding more IDs than the absolute judgment strategy (Wixted & Mickes, 2014). In laboratory experiments using multiple-lineups per participant, Meissner et al. (2005) used measures from signal detection theory to tease apart response bias and *discriminability* (the ability to distinguish between previously studied faces and new faces), and found that sequential lineups induced a more conservative response criterion but did not affect discriminability.

Mickes et al. (2012) argued that a particular kind of analysis, ROC (which arises from signal detection theory; see also Wixted & Mickes, 2015), is best suited to unambiguously answer the key question: which procedure better enables eyewitnesses to *discriminate* innocent from guilty suspects? To our knowledge, studies using ROC analysis have never reported a sequential superiority effect. Instead, they have found either no difference in discriminability between the two procedures or they have found a significant *simultaneous*

¹ Although technically inappropriate, as noted by Mickes, Moreland, Clark, and Wixted (2014), the common measure of discriminability, d' , can also be calculated from the data reported by Lindsay and Wells (1985), and it was higher for the sequential lineup (.95) than the simultaneous lineup (.38).

advantage (Andersen, Carlson, Carlson, & Gronlund, 2014; Carlson & Carlson, 2014; Dobolyi & Dodson, 2013; Gronlund et al., 2012; Meisters, Diedenhofen, & Musch, 2018; Mickes et al., 2012; Sučić, Tokić, & Ivešić, 2015).

Wixted and Mickes et al. (2014) proposed the diagnostic feature-detection theory to explain why simultaneous lineups are, if anything, superior to sequential lineups. Theoretically, compared to sequential lineups, simultaneous lineups help eyewitnesses to notice and to then discount non-diagnostic facial features (namely, the features that are common across the lineup members, such as race). By discounting non-diagnostic features, eyewitnesses are better able to focus attention on diagnostic features, enhancing their ability to tell the difference between innocent and guilty suspects (i.e., enhancing discriminability). This effect would show up as a higher ROC curve (or higher d') for simultaneous compared to sequential lineups.

In most eyewitness memory studies, subjects view one simulated crime and later complete a single lineup memory test, because this is similar to what happens in the real world. Such procedures yield one datum per subject and thus require large sample sizes in order to aggregate data to calculate a single discriminability statistic (e.g., area under the ROC or d') for each experimental condition. However, more traditional laboratory recognition experiments using simple stimuli can yield numerous observations per subject, allowing performance measures to be calculated for each subject, offering greater power and ability to generalize results across subjects; indeed, this was done by Meissner et al. (2005). Mansour, Beaudry, and Lindsay (2017) recently found that multiple-trial designs appear to have minimal effects on eyewitness identification performance relative to single-trial designs, and they therefore recommended that “Researchers should thus consider using multiple-trial designs for conducting eyewitness identification experiments” (p. 2235). Indeed, there is no obvious reason why the theoretical predictions made by the absolute/relative judgment theory and the diagnostic feature-detection theory cannot be tested using more traditional procedures even if the results would not have direct policy implications for police lineups.

The present study was designed to evaluate recognition memory performance yielded by simultaneous versus sequential lineups using basic experimental stimuli: word lists. In two experiments, subjects studied multiple lists of words, and recognition memory for those words was tested using six-item simultaneous or sequential word lineups. In order to simulate conditions used in face lineup studies, we used Deese-Roediger-McDermott lists (DRM; Deese, 1959; Roediger & McDermott, 1995), which consist of words that are all associated with a critical word that is missing from the list. For each list, subjects first studied part of the list (e.g., *bed, rest, tired, ...*); these words are analogous to perpetrators of a witnessed crime. In the subsequent lineup tests, a studied word

(e.g., *tired*) would act as the target, or guilty suspect. Unstudied words from the same DRM list (e.g., *yawn, nap, drowsy*) would act as fillers, or known innocents. The critical lure word, which was always unstudied (e.g., *sleep*), would act as an attractive innocent suspect.

Each subject was tested with four lineup types: target-present (TP), target-absent (TA), target-present-critical-lure (TPCL), and target-absent-critical-lure (TACL). We compared simultaneous versus sequential test formats in three ways. First, we used data from the TP and TA lineups (TP:TA), which are analogous to fair lineups. Second, we used data from the TPCL and TACL lineups (TPCL:TACL), to see how any effect of test format on discriminability is influenced by the inclusion of an attractive lure (the critical lure) to both target-present and target-absent lineups. This is analogous to a face lineup situation in which a known innocent *filler* face happens to have very similar features to the real perpetrator. Third, we used data from the TP and TACL lineups, focusing on false alarms specifically to the critical lure itself (TP:TACL_{cl}). This is analogous to an unfair or biased lineup situation in which the lineup either contains a guilty suspect (TP) or an unlucky innocent suspect (critical lure in TACL) who more closely resembles the actual perpetrator(s) than any of the fillers. For all three approaches (TP:TA, TPCL:TACL, and TP:TACL_{cl}), the question of interest was whether the simultaneous test format or the sequential test format yielded higher discriminability, as measured using both ROC and d' analysis.

Experiment 1

Method

Subjects Sixty-four undergraduates at Washington University in St. Louis (ages 18–25 years) participated in the study for course credit or payment. Sample size was based on what is typically used in DRM word-list experiments. DRM effects tend to be large (e.g., $d = 1.57$, Pardilla-Delgado & Payne, 2017; $d = 3.04$, Zhu, Chen, Loftus, Lin & Dong, 2013), and thus for example a sample size of only 42 is sufficient to detect a large between-subjects effect ($d \geq 0.80$) with 80% power. However, with regard to potential effect size for test format (simultaneous vs. sequential), there was very little relevant data available at the time this experiment was run (2013) to make a priori estimates. Although other effects found in lineup studies have tended to be small, those studies typically involve only one trial per subject. In our experiment, there were multiple trials per subject, which should thus reduce measurement error and increase effect size. In this experiment, achieved power was 88% for large between-subjects effect sizes ($d \geq 0.80$), and 50% for medium between-subjects effect sizes ($d =$

0.50). Upon discovering effect sizes smaller than anticipated in Experiment 1, we increased sample size, and thus power, in Experiment 2.

Design The experiment had two independent variables: test format, and lineup type. It used a 2×4 mixed design, with the two recognition test formats manipulated between subjects (simultaneous vs. sequential), and the four lineup types manipulated within subjects.² All lineups consisted of six items. The first two lineup types were: target-present (TP; one studied target word and five non-critical lures weakly related to that target) and target-absent (TA; six non-critical lures weakly related to the absent target). The remaining two lineup types replaced the weakest related lure (the one with the lowest backward associative strength to the target) with the critical lure, creating target-present-critical-lure (TPCL) and target-absent-critical-lure (TACL). As in real lineups, even the non-critical lures fit the criterion of being similar to (related to) the target item from the list, analogous to faces in a lineup matching the general verbal description given by an eyewitness. Nevertheless, the TACL lineup could be considered an unfair or biased lineup because the critical lure more closely resembles the actual perpetrator(s) than any of the fillers. Dependent variables were subjects' choices and confidence ratings on the lineups.

Materials Materials were 40 DRM lists taken from the norms provided by Roediger, Watson, McDermott, and Gallo (2001). We selected the 40 lists with the highest rates of false recognition of the critical lure word (*sleep*, in this example). Each list consisted of 15 words associated with the critical lure, which was not presented in the list. From each list we selected the ten words with the strongest backward associative strength to the critical lure to appear in the study phase, and we randomly selected one of the words from norm positions 4–8 (e.g., *tired*) to serve as the target word on the tests for all subjects. The remaining five DRM list items that were not studied were used as non-critical lures on the recognition tests. For the target-absent lineups, a sixth item was selected from the Nelson, McEvoy, and Schreiber (1998) norms to serve as the additional non-critical lure; it was weakly related to the critical lure and thus similar to the other non-critical lures. When the critical lure was included in a test lineup, we dropped the lure with the weakest backward associative strength to the critical lure. Figure 1 shows an example of an entire DRM list, the ten words that subjects studied, and the four different types of lineups that subjects saw at test.

TP target-present, *TA* target-absent, *TPCL* target-present with critical lure, *TACL* target-absent with critical lure

² This could also be considered a $2 \times 2 \times 2$ design: test format by target presence by critical lure presence. We combine the latter two variables under "lineup type" to simplify reporting of results.

Each subject studied and was tested on only 32 of the 40 DRM lists. Each DRM list was used equally often across the four lineup types, using a balanced Latin square to counterbalance order of conditions. For each block of eight lists for each subject, two were tested in each of the four lineup types.

Procedure This study received ethical approval from the Institutional Review Board of Washington University in St. Louis. The experiment was constructed using the Collector software (Garcia, n.d.) and displayed using the Google Chrome internet browser on 21.5-in. iMac computers. Subjects were instructed that they would be studying a total of 32 lists of ten words each for later recognition tests. They were told that there would be four study-test cycles – after studying eight lists, they would then be tested on those eight lists, and then repeat this cycle again for the remainder of the lists until all 32 lists were studied and tested.³

Subjects were randomly assigned to receive either all simultaneous or all sequential recognition tests, and were instructed about their test format. Both groups were told that in some tests, none of the words would be ones they studied, and that in such cases they should respond "No" to all items (sequential) or "Not present" (simultaneous). Subjects in the sequential condition were told that their first "Yes" response for each list would be the only one counted. Note, however, that these subjects still completed a trial for every one of the six words in a given test list regardless of their responses. Thus, the stopping rule was not implemented during the procedure, but we did implement it later during data analysis.

Subjects first completed a practice non-DRM list with an immediate test using the format they were assigned. After that, the procedure consisted of four study-test cycles. In the study phase of each cycle, words were presented in random order in 72-pt Times New Roman font, at a rate of 2 s per word with a 1-s inter-stimulus interval. A 10-s warning screen preceded presentation of each of the eight lists in the cycle. After studying all eight lists, subjects played the game Tetris for 2 min, and then began the recognition tests for that cycle. Lists were tested in the same order studied, to keep the retention intervals roughly consistent. Subjects were not told ahead of time how many words would be in the tests.

A 5-s warning slide appeared before each test. For the simultaneous test format, six words then appeared on the screen in a random order, arranged in two rows of three, with a "Not Present" button at the bottom. Subjects had 8 s to click one of the words or the "Not Present" button. If they did not respond in time, the program moved on to the next test. If they

³ See Supplemental Materials for complete instructions.

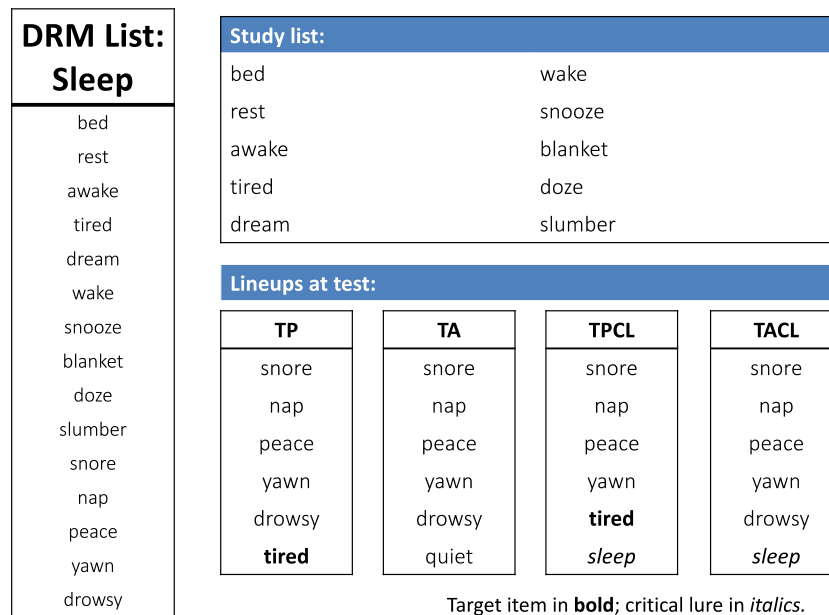


Fig. 1. Example DRM list, study list, and test lineup lists. Bold and italics are for illustrative purposes only; all words appeared in the same font in the actual tests. Actual word order was random for both study and test. Each subject was tested on only one lineup type per DRM list.

did respond in time, they then made a confidence rating using a slider that ranged from 20 to 100, where 20 was guessing and 100 was absolutely certain. Although chance was in fact 16.67%, we rounded up to 20 because the slider moved in increments of five.

For the sequential test format, subjects saw six words, one at a time, in a random order. For each word, subjects had 8 s to click either a “Yes” button or a “No” button. If they did not respond in time, the program moved on to the next word. If they did respond in time, they then made a confidence rating using a slider, just like in the simultaneous test format. The 8-s time limit was imposed for the practical purpose of keeping the entire procedure length under one hour.

This entire study-test cycle procedure was done four times for each subject, so that they ultimately studied and were tested on a total of 32 lists. Each lineup type (TP, TA, TPCL, TACL) was used for eight lists, and the order of lineup type was counterbalanced using a balanced Latin square. The entire procedure lasted approximately 60 min.

Results and discussion

An alpha level of .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. Standard deviations (*SDs*) 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 will first

report hit rates and false-alarm rates, followed by group ROC curves, and then individual *d'* scores.

Hit rates and false-alarm rates Table 1 shows the mean hit rates (correct ID rates) and false-alarm rates (false ID rates) by test format (simultaneous vs. sequential) and lineup type. In some eyewitness studies, a stopping rule is used for sequential lineups, such that the remainder of a lineup is not presented once the subject responds “yes” on any trial. Steblay et al. (2011, p. 117) reported that a majority of eyewitness research teams did not use such a rule. In the current experiment, although subjects in the sequential condition were told that only their first “yes” would count, they still completed a trial for every one of the six words in a given test list regardless of their responses. Thus, the stopping rule was not executed during the procedure, but for data analysis we simulated it by disregarding any responses after the first “yes” response in each sequential lineup. We will report results first with the stopping rule, then without it.

In Table 1, the hit rate (correct ID rate) is the proportion of target-present lineups (either TP or TPCL) in which the subject chose the single target word that had been previously studied. The table also shows several different calculations of false-alarm rate (false ID rate). The critical lure false ID rate is the proportion of lineups with a critical lure (either TPCL or TACL) in which the subject chose the critical lure. The non-critical lure false ID rate is the proportion of lineups in which the subject chose any of the non-critical lures. Note that the number of non-critical lures varied across lineup type (TP: 5, TPCL: 4, TA: 6, TACL: 5). Thus, we also report the non-critical lure false ID rate adjusted by dividing by the

Table 1. Mean correct ID (hit) rates and false ID (false-alarm) rates in Experiment 1

Test format × lineup type	Hit rate	False-alarm rate			Lineup rejection rate
		Critical lure	Non-critical lures	Non-critical lures (adjusted)	
Simultaneous					
TP	.56 (.25)		.22 (.23)	.04 (.05)	.21 (.20)
TA			.36 (.26)	.06 (.04)	.59 (.25)
TPCL	.57 (.26)	.15 (.17)	.13 (.17)	.03 (.04)	.16 (.13)
TACL		.28 (.24)	.24 (.17)	.05 (.03)	.44 (.25)
Sequential					
TP	.56 (.26)		.28 (.24)	.06 (.05)	.17 (.16)
TA			.45 (.33)	.07 (.05)	.55 (.33)
TPCL	.38 (.28)	.26 (.19)	.22 (.21)	.05 (.05)	.14 (.19)
TACL		.29 (.21)	.29 (.28)	.06 (.06)	.42 (.32)

Note. Standard deviations in parentheses. Adjusted false-alarm rates for non-critical lures are divided by the number of such lures in the given lineup type (TP: 5, TA: 6, TPCL: 4, TACL: 5). Stopping rule (first yes) was in effect for sequential condition. Overall false-alarm rates in TPCL and TACL can be computed by adding false-alarm rates for critical and non-critical lures

number of non-critical lures in that lineup type. This adjustment to false-alarm rate is typically done in eyewitness lineup studies when target-absent lineups do not contain a designated-innocent suspect. The adjustment thus estimates the probability of a subject choosing any one particular non-critical lure.

Finally, Table 1 also shows the mean lineup rejection rate, which is the proportion of lineups in which the subject did not choose any of the items; such responses constituted misses in target-present lineups and correct rejections in target-absent lineups. As expected, lineup rejections were more likely for target-absent lineups than target-present lineups for both the simultaneous test format, $t(31) = 7.39$, $p < .001$, $d = 1.73$, and the sequential test format, $t(31) = 6.86$, $p < .001$, $d = 1.34$.

TP target-present, *TA* target-absent, *TPCL* target-present with critical lure, *TACL* target-absent with critical lure

What was the effect of test format (simultaneous vs. sequential)? Consider first the TP and TA lineups, which did not contain critical lures. The hit rate for simultaneous TP lineups (correct ID rate = .56) did not significantly differ from the hit rate for sequential TP lineups (correct ID rate = .56), $t(62) = 0.01$, $p = .990$, $d < 0.01$, and the adjusted false-alarm rates to non-critical lures for TA lineups were also similar across test formats (.06 for simultaneous, and .07 for sequential), $t(62) = 1.11$, $p = .273$, $d = 0.28$. The hit rate for simultaneous TPCL lineups (correct ID rate = .57) was significantly higher than the hit rate for sequential TPCL lineups (correct ID rate = .38), $t(62) = 2.70$, $p = .009$, $d = 0.67$, whereas the adjusted false-alarm rates to non-critical lures for TACL lineups were similar across test formats (.05 for simultaneous, and .06 for sequential), $t(62) = 0.91$, $p = .366$, $d = 0.23$.

As expected, due to the nature of DRM lists, in target-absent lineups that contained a critical lure, subjects were more likely to false alarm to the critical lure than to any one

particular non-critical lure (.28 vs. .05 for TACL simultaneous, and .29 vs. .06 for TACL sequential lineups). That effect accounts for the reduction in the non-critical false-alarm rates for both simultaneous and sequential test formats when the critical lure was included in the lineups.

ROC analysis We next measured discriminability using ROC analysis, focusing on three comparisons. First, we considered the ability to distinguish between targets and lures in TP:TA lineups, analogous to a fair lineup situation. Second, we considered the ability to distinguish between targets and non-critical lures in TPCL:TACL lineups, which addresses the effect of having an attractive filler (critical lure) added to both the TP and TA lineups. Third, we considered the ability to distinguish between targets and critical lures in TP:TACL_cl lineups, counting only false alarms to the critical lure, analogous to an unfair or biased lineup situation in which the lineup either contains a guilty suspect (TP) or an unlucky innocent suspect (critical lure in TACL) who more closely resembles the actual perpetrator(s) than any of the fillers. Because the data were too noisy to construct ROCs at the individual level, we combined data from all subjects to construct group ROCs, using the pROC statistical package for R (Robin et al., 2011), as described in the tutorial provided in the supplemental materials of Gronlund, Wixted, and Mickes (2014).⁴ Figure 2 shows group ROCs for simultaneous versus sequential test formats. Figure 2A presents data from the fair lineups (TP:TA); Fig. 2B presents data from lineups with an attractive lure (TPCL:TACL, counting only non-critical false alarms); and Fig. 2C presents data from unfair target-absent lineups (TP:TACL_cl, with the critical lure serving as the innocent suspect). For ROC figures and analyses we did not bother with

⁴ See Supplemental Materials for details on data formatting.

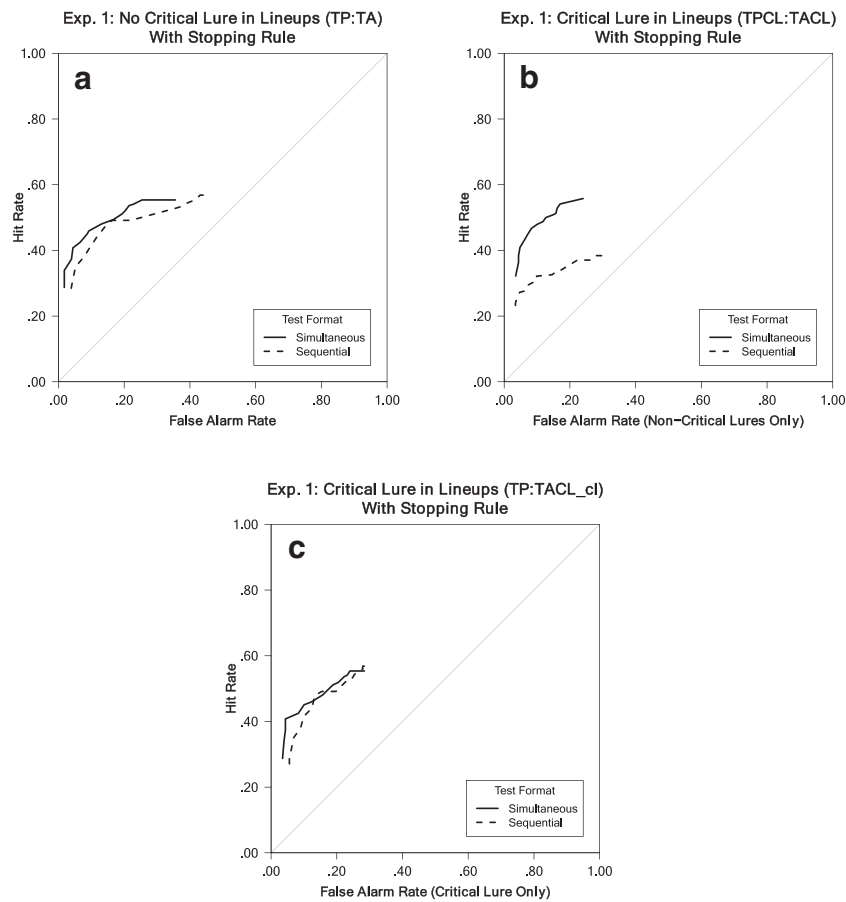


Fig. 2. Mean receiver operating characteristics (ROCs) for simultaneous and sequential test formats in Experiment 1, with stopping rule (first yes) used for the sequential test format. ROCs were constructed using (A) lineups without the critical lure (TP:TA), (B) lineups with the critical lure (TPCL:TACL), and (C) lineups with the critical lure in target-absent lineups only and with the critical lure serving as the “innocent suspect” (TP:TACL_cl). The solid diagonal line represents chance performance

any adjustments to false-alarm rates (i.e., dividing by number of lures), because such a linear transformation would only change the labeling of the x-axis in the figures (Wixted & Mickes, 2015, point 2), and would not influence any inferential statistical tests.

It is plain to see that when critical lures were included in the lineups (Fig. 2B), the simultaneous test format yielded superior discrimination compared to the sequential test format; but any difference between test formats is less clear in the other two panels. Still, we see no hint of a sequential superiority effect. In order to statistically compare the partial area under the curve (pAUC) across test formats, it is necessary to first determine the range of false-alarm rates to use for calculation. Typically, the area under the ROCs for lineups is defined by a false-alarm rate range extending from 0 to the maximum false-alarm rate of one of the two conditions being compared. In our case, we could use the maximum FAR yielded by either the simultaneous test format (which tended to be smaller) or the sequential test format (which tended to be larger). We used both approaches, and the results were very similar either way. For each comparison, the test statistic provided by the pROC

package is $D = (pAUC_{sim} - pAUC_{seq})/s$ where s is the standard error of differences under the null hypothesis, which was estimated using non-parametric bootstrapping (resampling with replacement, 10,000 replicates). The p -value was obtained by comparing the test statistic to a standard normal distribution. Table 2 shows the results of these analyses.

Observing the rows in Table 2 where the stopping rule was used for the sequential test format, we see that when there were no critical lures in the lineups (i.e., TP:TA), there was no significant difference between the simultaneous and sequential test formats (Fig. 2A). However, when critical lures were used in the lineups (i.e., TPCL:TACL), we see that pAUC was statistically significantly greater for the simultaneous test format compared to the sequential test format. This indicates that, when critical lures were present, subjects in the simultaneous condition were better able to discriminate between targets and lures than subjects in the sequential condition, as is plainly visible in Fig. 2B. Finally, we also see that when critical lures were in the lineup and only identification of the critical lure was considered a false alarm (i.e., TP:TACL_cl), there was

Table 2. Comparison of partial area under the curve (pAUC) for group ROCs in Experiment 1

Lineup type	Stopping rule	Max FAR used for pAUC	Simultaneous pAUC	Sequential pAUC	<i>D</i>	<i>p</i>
No critical lure in lineups (TP:TA)						
Stopping rule						
	Smaller		.17	.15	1.00	.315
	Larger		.22	.20	0.98	.326
No stopping rule						
	Smaller		.17	.14	2.10	.036
	Larger		.22	.18	2.04	.042
Critical lure in lineups (TPCL:TACL)						
Stopping rule						
	Smaller		.11	.07	3.33	< .001
	Larger		.14	.09	3.59	< .001
No stopping rule						
	Smaller		.11	.06	4.49	< .001
	Larger		.12	.07	4.55	< .001
Critical lure in lineups (TP:TACL_cl)						
Stopping rule						
	Smaller		.12	.11	0.73	.466
	Larger		.13	.12	0.73	.466
No stopping rule						
	Smaller		.13	.10	2.05	.040
	Larger		.14	.11	1.97	.048

Note. $D = (pAUC_{sim} - pAUC_{seq})/s$, where s is the standard error of differences under the null hypothesis, which was estimated using non-parametric bootstrapping (resampling with replacement, 10,000 replicates); pAUC is calculated from 0 to the max FAR, which differed across lineup types, thus making it inappropriate to compare these pAUC values between lineup types. Stopping rule (first yes) was implemented or not during data analysis

no significant difference between the simultaneous and sequential test formats (Fig. 2C). Again, this is only considering results with the stopping rule in effect.

Individual d' analysis In addition to these group ROC analyses, we also computed d' scores from the individual subject hit and false-alarm rate data, with the commonly used adjustment of adding or subtracting $\frac{1}{2}$ of a hit or false alarm to avoid rates of 0 or 1 (Macmillan & Creelman, 2004, p. 8). Mickes et al. (2014) showed that although d' was designed for an old/new recognition procedure, it nevertheless does a reasonably good job in measuring discriminability for lineups. The results largely corresponded to the ROC analyses. For the TP:TA d' values, calculated using the hit rate from TP and the adjusted false-alarm rate from TA, the mean was not significantly different for the simultaneous test format (1.70 , $SD = 0.87$) versus the sequential test format (1.66 , $SD = 1.07$), $t(62) = 0.18$, $p = .856$, $d = 0.05$. For the TPCL:TACL d' values, calculated using the hit rate from TPCL and the adjusted non-critical false-alarm rate from TACL, the mean was significantly higher for the simultaneous test format (1.80 , $SD = 0.83$) versus the sequential test format (1.15 , $SD = 0.94$), $t(62) = 2.84$, $p = .006$, $d = 0.71$. Finally, we consider the TP:TACL_cl d' values, calculated using the hit rate from TP and the critical

false-alarm rate from TACL, which represent a biased lineup situation. The mean was not significantly different for the simultaneous test format (0.80 , $SD = 1.14$) versus the sequential test format (0.80 , $SD = 1.07$), $t(62) = 0.01$, $p = .991$, $d < 0.01$. Note that this test is not independent of the main tests we report using the TP:TA and TPCL:TACL lineup combinations, but we include it for completeness.

Removing the stopping rule The results reported above for the lineups featuring critical lures offer some support for predictions made by the diagnostic feature-detection theory (Wixted & Mickes, 2014). That is, at least for the TPCL:TACL lineup combination, the simultaneous test format was associated with higher discriminability than the sequential test format. However, a stopping rule was used in the analysis of the sequential test format. As shown by Rotello and Chen (2016), the stopping rule imposes an artificial constraint on discriminative performance, one that increases as responding becomes more liberal. That is, as responding becomes more liberal, measured discriminative performance (empirical discriminability) becomes worse even though underlying discriminability (the degree to which the internal memory signals associated with targets and lures overlap) remains the same. In effect, the stopping rule makes underlying discriminability appear to

be worse than it really is. Conceivably, the entire difference in discriminability between simultaneous and sequential lineup formats (Table 2, Fig. 2) may result from this artificial constraint, in which case the data would not lend any support to the diagnostic feature-detection theory after all.

We therefore reanalyzed the sequential data after removing the stopping rule. In 34% of the lineups, subjects made more than one “yes” response. Our procedure for removing the stopping rule was as follows: for each lineup and each subject, if there was more than one “yes” response, then instead of using the first “yes” response made by the subject, we used the “yes” response made with the highest level of confidence as the official ID decision for that lineup and subject. In case of a tie (e.g., a subject said “yes” with 100% confidence to both the second and the fifth items in a lineup), we used the last “yes” on the assumption that it overrode previous IDs. Such ties occurred in only 8% of all lineups.

The ROC curves are shown in Fig. 3, and a simultaneous advantage is now apparent in all three comparisons. As shown in the “no stopping rule” rows of Table 2, without the stopping rule, the greater pAUC for simultaneous versus sequential test

format became statistically significant in the absence of critical lures (TP:TA, Fig. 3A) and remained so in their presence (TPCL:TACL, Fig. 3B). For the TP:TACL_cl ROC curves (Fig. 3C), the effect of test format also became significant in favor of simultaneous lineups, whereas it had not been significant with the stopping rule.

The pattern of results based on group ROC analysis without the stopping rule was also apparent in the d' scores computed from individual subject hit and false-alarm rate data, though significant differences obtained using ROC analysis were not always significant in the analysis of d' scores. For the TP:TA individual d' analysis, and in agreement with the corresponding group ROC analysis, the mean d' for the simultaneous test format (1.70, $SD = 0.87$) was higher than the mean d' for the sequential test format (1.54, $SD = 1.10$). However, in contrast to the ROC analysis, the difference was not significantly different, $t(62) = 0.66, p = .511, d = 0.17$. For the TPCL:TACL individual d' analysis, and in full agreement with the group ROC analysis, the mean d' was significantly higher for the simultaneous test format (1.80, $SD = 0.83$) versus the sequential test format (1.05, $SD =$

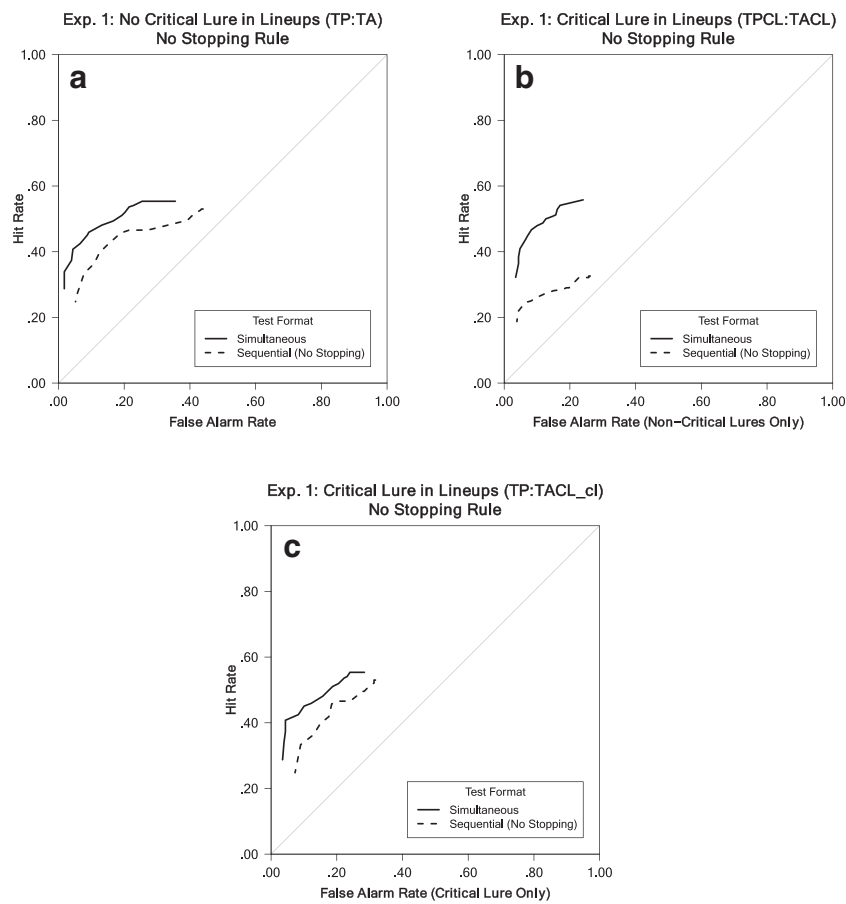


Fig. 3. Mean receiver operating characteristics (ROCs) for simultaneous and sequential test formats in Experiment 1, with no stopping rule used for the sequential test format. ROCs were constructed using (A) lineups without the critical lure (TP:TA), (B) lineups with the critical lure

(TPCL:TACL), and (C) lineups with the critical lure in target-absent lineups only and with the critical lure serving as the “innocent suspect” (TP:TACL_cl). The solid diagonal line represents chance performance

0.97), $t(62) = 3.26$, $p = .002$, $d = 0.82$. Finally, for the TP:TACL_cl individual d' analysis, and again in agreement with the corresponding group ROC analysis, the mean d' for the simultaneous test format (0.80, $SD = 1.14$) was higher than the mean d' for the sequential test format (0.68, $SD = 1.11$). However, unlike in the ROC analysis, the difference was not significantly different, $t(62) = 0.41$, $p = .681$, $d = 0.10$.

The effect of removing the stopping rule (which, if anything, lowered discriminability for sequential lineups) is the opposite of what we predicted based on past research. Usually, the removal of the stopping rule *elevates* the sequential ROC (Rotello & Chen, 2016; Wilson, Donnelly, Christenfeld, & Wixted, 2019). However, in Experiment 1, for reasons unknown, it seems that subjects in the sequential condition who happened to correctly choose the target as their first decision in TP and TPCL lineups tended to choose a subsequent lure with equal or higher confidence, thereby decreasing the hit rate when the stopping rule was removed. Whatever the reason for this unexpected result, the overall pattern of results, showing a decisive lack of advantage for the sequential test format, does not appear to be due to the stopping rule. In fact, in analyses without the stopping rule, the simultaneous test format appears to be superior to the sequential test format (definitively so according to the group ROC analysis and less definitively so according to an analysis of individual d' scores).

Limitations One limitation of Experiment 1 was that subjects did not always respond within the 8-s response deadline. Therefore, data were missing for some trials (Simultaneous: 6% in TP, 8% in TA, 2% TPCL, 8% in TACL; Sequential: 3% in TP, 4% in TA, 7% TPCL, 3% in TACL). Furthermore, the 8-s response deadline may have served to bias the results in favor of the sequential format. For the simultaneous lineups, the deadline meant that subjects had only 8 s to consider six words, which means that they had only 1.33 s of decision-making time per word. For the sequential lineups, subjects had a full 8 s of decision-making time per word. Any potential effect of the difference in decision-making time imposed by the response deadline was eliminated in Experiment 2, which used no response deadlines. Finally, due to a programming error, target items for two particular DRM lists were not properly presented in TP and TPCL lineups, so data from those lists were excluded from analysis. This too was fixed in Experiment 2.

Some eyewitness lineup studies comparing simultaneous and sequential procedures have used a “backloading” technique, by which subjects in the sequential condition do not know exactly how many photos they will see because they are not informed of the number or are told there will be more photos than there really are. This has been found to encourage a more conservative response bias throughout the sequential lineup (Carlson, Carlson, Weatherford, Tucker, & Bednarz,

2016; Horry et al., 2012). One limitation of the use of multiple lineups per subject, as in the current study, is that backloading cannot be used, because subjects would know how long the lineups are after the first few lineups.

Experiment 2

The results of Experiment 1 showed an advantage of the simultaneous test format over the sequential test format, but only when lineups contained critical lures. However, because we imposed an 8-s response deadline, subjects did not respond on some trials in all the conditions, and the amount of decision time per word favored the sequential test format. Experiment 2 was designed to replicate the results of Experiment 1 while eliminating the time limit and also correcting a programming error so that targets were properly presented for all DRM lists. All other methods were the same as in Experiment 1.

Method

Subjects Ninety-six undergraduates at Washington University in St. Louis (ages 18–25 years) participated in the study for course credit or payment. In this experiment, achieved power was 97% for large between-subjects effect sizes ($d \geq 0.80$), and 68% for medium between-subjects effect sizes ($d = 0.50$).

Design and materials The design and materials were the same as in Experiment 1.

Procedure The procedure was the same as in Experiment 1, with the exception that subjects were required to respond on each test trial before they moved on to the next trial, with no time limit. Due to a programming error, test instructions for the simultaneous condition still mentioned an 8-s limit, although no such limit was actually implemented, and subjects in that condition frequently took longer than 8 s to respond (mean of 15% of trials across subjects).

Results and discussion

Hit rates and false-alarm rates Table 3 shows the mean hit rates (correct ID rates) and false-alarm rates (false ID rates) by test format (simultaneous vs. sequential) and lineup type for Experiment 2. As expected, and as in Experiment 1, lineup rejections were more likely for target-absent lineups than target-present lineups for both the simultaneous test format, $t(47) = 9.93$, $p < .001$, $d = 2.02$, and the sequential test format, $t(47) = 8.73$, $p < .001$, $d = 1.34$.

What was the effect of test format in the absence of critical lures? The hit rate for simultaneous TP lineups (correct ID rate = .62) was significantly higher than the hit rate for sequential TP lineups (correct ID rate = .48), $t(94) = 2.87$, $p = .005$, $d =$

Table 3. Mean correct ID (hit) rates and false ID (false-alarm) rates in Experiment 2

Test format × lineup type	Hit rate	False-alarm rate			Lineup rejection rate
		Critical lure	Non-critical lures	Non-critical lures (adjusted)	
Simultaneous					
TP	.62 (.26)		.21 (.19)	.04 (.04)	.17 (.18)
TA			.45 (.24)	.07 (.04)	.55 (.24)
TPCL	.61 (.25)	.15 (.14)	.13 (.16)	.03 (.04)	.11 (.12)
TACL		.32 (.20)	.27 (.22)	.05 (.04)	.41 (.24)
Sequential					
TP	.48 (.21)		.30 (.24)	.06 (.05)	.22 (.17)
TA			.50 (.27)	.08 (.05)	.50 (.27)
TPCL	.45 (.22)	.18 (.13)	.23 (.19)	.06 (.05)	.14 (.16)
TACL		.32 (.19)	.32 (.20)	.06 (.04)	.36 (.23)

Note. Standard deviations in parentheses. Adjusted false-alarm rates for non-critical lures are divided by the number of such lures in the given lineup type (TP: 5, TA: 6, TPCL: 4, TACL: 5). Stopping rule (first yes) was in effect for sequential condition. Overall false-alarm rates in TPCL and TACL can be computed by adding false-alarm rates for critical lure and non-critical lures. *TP* target-present, *TA* target-absent, *TPCL* target-present with critical lure, *TACL* target-absent with critical lure

0.59, whereas the adjusted false-alarm rates to non-critical lures for TA lineups were again similar across test formats (.07 for simultaneous, and .08 for sequential), $t(94) = 1.07$, $p = .287$, $d = 0.22$. The pattern was similar for the TPCL and TACL lineups, which did contain critical lures. The hit rate for simultaneous TPCL lineups (correct ID rate = .61) was significantly higher than the hit rate for sequential TPCL lineups (correct ID rate = .45), $t(94) = 3.42$, $p = .001$, $d = 0.70$, whereas the adjusted false-alarm rates to non-critical lures for TACL lineups were again similar across test formats (.05 for simultaneous, and .06 for sequential), $t(94) = 1.21$, $p = .228$, $d = 0.25$.

As in Experiment 1, in lineups that contained a critical lure subjects were more likely to false alarm to the critical lure than to any one particular non-critical lure (.32 vs. .05 for TACL simultaneous, and .32 vs. .06 for TACL sequential). That effect accounts for the reduction in the non-critical false-alarm rates for both simultaneous and sequential lineups when the critical lure was included in the lineups. Overall, hit rates were higher for simultaneous versus sequential lineups, and false-alarm rates did not significantly differ.

ROC analysis ROC curves were constructed and analyzed in the same way described in Experiment 1. Figure 4 shows group ROCs for simultaneous versus sequential test formats, and as a function of whether or not a critical lure was present in the lineups. Figure 4A presents data from the fair lineups (TP:TA); Fig. 4B presents data from lineups with an attractive lure (TPCL:TACL); and Fig. 4C presents data from unfair target-absent lineups (TP:TACL_cl, with the critical lure serving as the innocent suspect). Table 4 shows comparisons of pAUC for simultaneous versus sequential test formats, conducted just as described in Experiment 1. Regardless of which

right-hand cutoff was used, and whether or not critical lures were present in the lineups, and whether non-critical or critical lures were counted as false alarms, pAUC was larger for the simultaneous versus sequential test format in every situation. This difference reached statistical significance in all cases except TPCL:TACL with no stopping rule, for which the difference was only marginally statistically significant.

Individual d' analysis As in Experiment 1, we also analyzed d' computed from the individual subject data. In the absence of critical lures (TP:TA), mean d' was significantly greater for the simultaneous test format (1.82, $SD = 0.89$) versus the sequential test format (1.37, $SD = 0.76$), $t(94) = 2.65$, $p = .009$, $d = 0.54$. In the presence of critical lures (TPCL:TACL), mean d' was also significantly greater for the simultaneous test format (1.93, $SD = 0.84$) versus the sequential test format (1.40, $SD = 0.75$), $t(94) = 3.22$, $p = .002$, $d = 0.66$. For the TP:TACL_cl d' values, the mean was also significantly greater for the simultaneous test format (0.89, $SD = 0.97$) versus the sequential test format (0.47, $SD = 0.83$), $t(94) = 2.27$, $p = .026$, $d = 0.46$. This last result is in contrast to the view in the eyewitness literature that a sequential test format particularly protects against biased lineups (Lindsay et al., 1991). Thus, results from Experiment 2 showed an unqualified advantage for the simultaneous test format in terms of discriminability as measured by d' .

Removing the stopping rule As in Experiment 1, we also recalculated all analyses without the stopping rule for the sequential test format. The ROC curves are shown in Fig. 5, and a simultaneous advantage is still apparent in all three

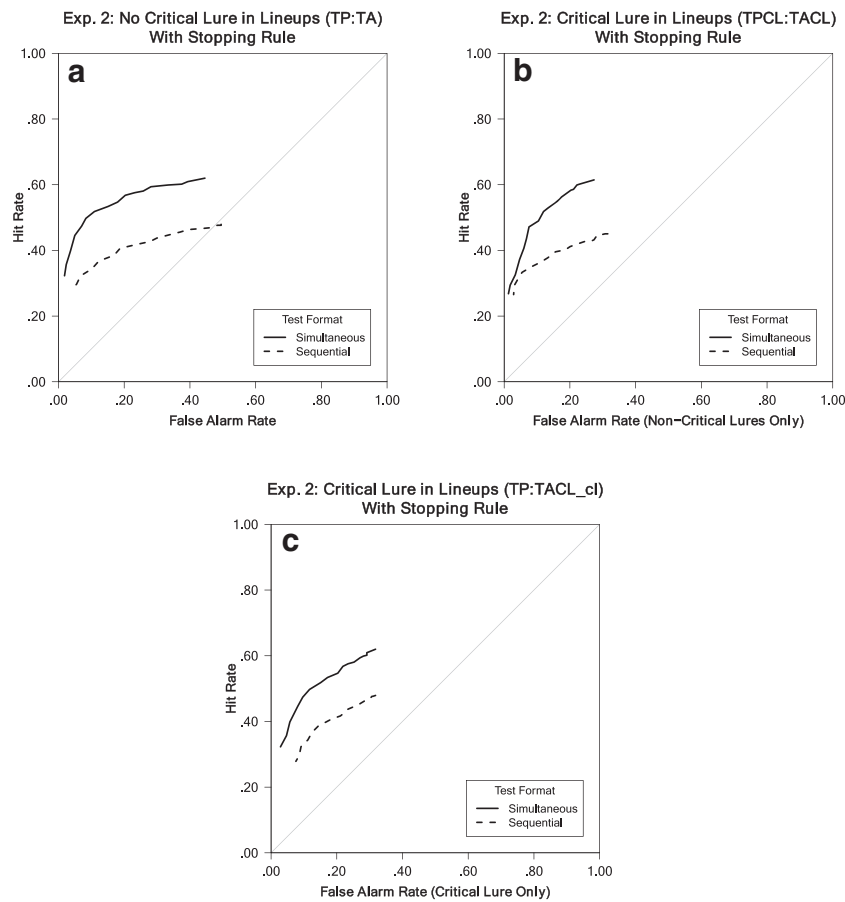


Fig. 4 Mean receiver operating characteristics (ROCs) for simultaneous and sequential test formats in Experiment 2, with stopping rule (first yes) used for the sequential test format. ROCs were constructed using (A) lineups without the critical lure (TP:TA), (B) lineups with the critical lure

(TPCL:TACL), and (C) lineups with the critical lure in target-absent lineups only and with the critical lure serving as the “innocent suspect” (TP:TACL_cl). The solid diagonal line represents chance performance

comparisons, though not as large as with the stopping rule (Fig. 4). As shown in the “no stopping rule” rows of Table 4, the pAUC values for the sequential test format in this case increased once we removed the artificial constraint imposed by the stopping rule. The trend differs from what we observed in Experiment 1 but is now in the direction predicted by signal detection accounts of sequential lineup performance (Rotello & Chen, 2016; Wilson et al., 2019). Nevertheless, removing the stopping rule did not change the pattern of results. The greater pAUC for simultaneous versus sequential test format remained statistically significant in all three cases: TP:TA, TPCL:TACL, and TP:TACL_cl (Table 4). Thus, based on pAUC, the overall pattern of results, showing an advantage for the simultaneous test format, does not appear to be an artifact of the stopping rule.

Similarly, all of the d' trends still favored the simultaneous procedure, but the effect sizes were reduced relative to the analyses that used the stopping rule and the differences were no longer significant. For TP:TA, $M_{sim} = 1.82$, $SD_{sim} = 0.89$, $M_{seq} = 1.66$, $SD_{seq} = 0.71$, $t(94) = 0.90$, $p = .369$, $d = 0.18$. For TPCL:TACL, $M_{sim} = 1.93$, $SD_{sim} =$

0.84 , $M_{seq} = 1.68$, $SD_{seq} = 0.71$, $t(94) = 1.51$, $p = .133$, $d = 0.31$. For TP:TACL_cl, $M_{sim} = 0.89$, $SD_{sim} = 0.97$, $M_{seq} = 0.64$, $SD_{seq} = 0.81$, $t(94) = 1.41$, $p = .162$, $d = 0.29$.

General discussion

Across two experiments we used DRM word lists to simulate eyewitness lineup procedures. Using ROC analysis in a multi-trial, word-list lineup paradigm, we found that simultaneous lineups generally yielded superior discriminability compared to sequential lineups (with the exception of TP:TA in Experiment 1 when using the simulated stopping rule, in which no difference was found). This pattern mirrors the literature using face lineups, which almost always finds evidence favoring the simultaneous procedure, though such trend is not always significant (Andersen et al., 2014; Carlson & Carlson, 2014; Dobolyi & Dodson, 2013; Gronlund et al., 2012; Mickes et al., 2012; Terrell, Baggett, Dasse, & Malavanti, 2017). Indeed, we know of only one study using ROC analysis in which the trend favored the sequential

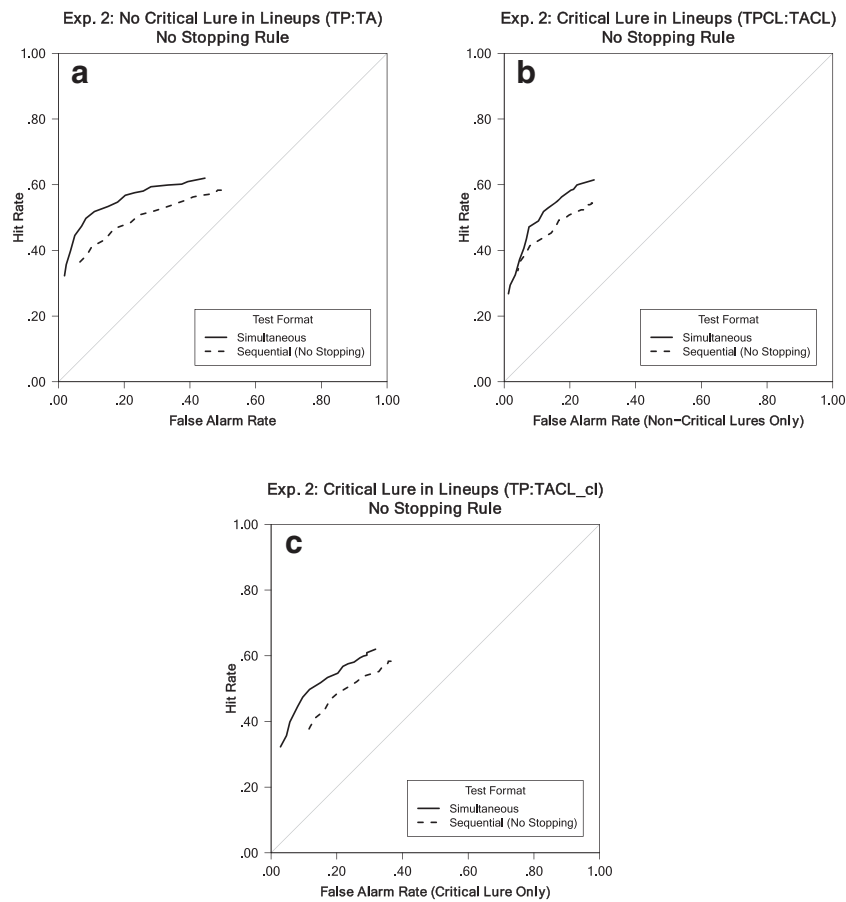


Fig. 5. Mean receiver operating characteristics (ROCs) for simultaneous and sequential test formats in Experiment 2, with no stopping rule used for the sequential test format. ROCs were constructed using (A) lineups without the critical lure (TP:TA), (B) lineups with the critical lure

(TPCL:TACL), and (C) lineups with the critical lure in target-absent lineups only and with the critical lure serving as the “innocent suspect” (TP:TACL_cl). The solid diagonal line represents chance performance

procedure (Meisters et al., 2018). Our findings are the first to generalize the apparent simultaneous superiority effect from the perceptual/face domain to the semantic/word domain.

When prior researchers have argued for the superiority of sequential lineup procedures, one of the key pieces of evidence has been lower false identification rates found in sequential versus simultaneous lineups (cf. Steblay et al., 2011, p. 106). That is, a sequential procedure appears to yield a more conservative response bias, with eyewitnesses less likely to choose anyone from the lineup (Palmer & Brewer, 2012). In the current study we did not find lower false-alarm rates for the sequential test format as compared to the simultaneous test format. However, several recent studies using traditional eyewitness lineup materials (i.e., faces) have also found no conservative response bias for the sequential procedure (Carlson & Carlson, 2014; Andersen et al., 2014). Furthermore, if our goal is to determine which procedure best facilitates discrimination of guilty from innocent suspects, be they words or faces, then response bias is beside the point. Response bias can be readily shifted by factors such as instructions (Clark, 2005; Meissner et al., 2005; Rotello &

Macmillan, 2007). The key issue is discriminability, as measured using ROCs and d' .

One curious matter is why implementing the stopping rule (as compared to no stopping rule) appeared to boost performance for the sequential test format in Experiment 1 but reduce performance in Experiment 2. Only the latter effect is anticipated by any basic signal detection model of sequential lineup performance (Rotello & Chen, 2016; Wilson et al., 2019). We do not have a ready explanation for this difference. The main difference between the two experiments was the presence (Experiment 1) or absence (Experiment 2) of an 8-s time limit for responses. However, excluding the small number of trials for which subjects failed to respond in time in Experiment 1, the median response time for sequential trials was very similar across experiments: 1.28 s in Experiment 1, and 1.26 s in Experiment 2. Nevertheless, perhaps merely being aware of a time limit affected subjects’ absolute judgments in Experiment 1’s sequential format in such a way that their performance was best when measured by their first response (stopping rule) rather than their highest confidence response (no stopping rule). In any event, even with the

Table 4. Comparison of partial area under the curve (pAUC) for group ROCs in Experiment 2

Lineup type	Stopping rule	Max FAR used for pAUC	Simultaneous pAUC	Sequential pAUC	<i>D</i>	<i>p</i>
No critical lure in lineups (TP:TA)						
Stopping rule						
		Smaller	.24	.17	4.44	< .001
		Larger	.27	.19	4.52	< .001
No stopping rule						
		Smaller	.24	.20	2.55	.011
		Larger	.27	.23	2.46	.014
Critical lure in lineups (TPCL:TACL)						
Stopping rule						
		Smaller	.13	.10	3.71	< .001
		Larger	.16	.12	3.92	< .001
No stopping rule						
		Smaller	.13	.12	1.81	.070
		Larger	.13	.11	1.82	.069
Critical lure in lineups (TP:TACL_cl)						
Stopping rule						
		Smaller	.15	.11	3.92	< .001
		Larger	.15	.11	3.92	< .001
No stopping rule						
		Smaller	.15	.12	2.84	.004
		Larger	.18	.15	2.73	.006

Note. $D = (pAUC_{sim} - pAUC_{seq})/s$ where s is the standard error of differences under the null hypothesis, which was estimated using non-parametric bootstrapping (resampling with replacement, 10,000 replicates); pAUC is calculated from 0 to the max FAR, which differed across lineup types, thus making it inappropriate to compare these pAUC values between lineup types. Stopping rule (first yes) was implemented or not during data analysis

unexpected boost afforded by the stopping rule in Experiment 1, the group ROC curve for the sequential test format was never higher than the curve for the simultaneous test format. Thus, our main pattern of results stands: we found no superiority of the sequential test format over the simultaneous test format, but rather in most cases found quite the opposite: simultaneous test formats yielded the superior discriminability.

As noted above, our finding of simultaneous format superiority corresponds to previous results obtained using single-trial mock-crime lineup paradigms involving faces (e.g., Carlson & Carlson, 2014; Dobolyi & Dodson, 2013; Mickes et al., 2012). Because we used word lists, our results do not have direct policy implications, but they do have theoretical implications. The sequential lineup format reduces discriminability, and it appears to do so for two distinct reasons. First, when the stopping rule is used, discriminative performance is generally found to be artificially impaired by preventing subjects from ever seeing the suspect (Rotello & Chen, 2016; Wilson et al., 2019). Consider, for example, the extreme situation in which subjects in an eyewitness experiment are maximally liberal, in which case they would always choose the first face in the six-person sequential lineup. Because the guilty suspect is in the first position 1/6 of the time, the hit

rate would be 1/6. However, the innocent suspect is also in the first position 1/6 of the time, so the false-alarm rate would also be 1/6. In that case, empirical discriminability (pAUC or measured d') would indicate zero ability to discriminate targets from lures, and this would be true even for subjects whose underlying d' (difference between their internal memory signals for targets versus lures) is much greater than 0. The stopping rule does serve to reduce the false-alarm rate (a positive feature), but the cost of impairing discriminability can be high. Thus, it is perhaps fortunate that when police departments have adopted the sequential procedure, they rarely, if ever, use a stopping rule (Stebly et al., 2011), even though some of the research that led police to change to sequential presentation lineups was based on using a stopping rule.

Aside from the artificial constraint imposed by the stopping rule, a second factor also theoretically reduces discriminability at the psychological level when a sequential format is used. Wixted and Mickes et al. (2014) proposed that the simultaneous presentation of faces in a lineup allows eyewitnesses to detect and discount common (and therefore non-diagnostic) facial features. Attaching less weight to non-diagnostic features (i.e., features shared by guilty suspects, innocent suspects, and the fillers) should enhance discriminability. Essentially, this theory holds that the memory-strength signal

in a lineup is determined by the degree to which one face stands out from the other faces in terms of how well it matches the memory of the perpetrator. By design, all of the faces in the lineup match the memory trace of the perpetrator to some degree, because faces are included in the lineup only if they match the general description of the perpetrator. Thus, focusing on the *difference* in the memory signal generated by one face versus the average memory signal generated by all the faces in the lineup subtracts away the contribution of the common features to that signal (Wixted, Vul, Mickes, & Wilson, 2018). Doing so yields an operative memory signal that is more diagnostic of guilt (i.e., one that is based on non-shared memory-matching features uniquely associated with the perpetrator) than would otherwise be the case. Analogously, a within-subjects experimental design is more powerful than a between-subjects design because the former subtracts out measurement error that might otherwise mask a genuine effect.

Our data suggest that the same theoretical interpretation may also apply to semantic features when lists of words are used. In a word lineup, it is not physical features that are shared by design; instead, in our lists, semantic features were shared. To the degree that those shared semantic features contribute to the memory signal, they would add noise, thereby making it harder to discriminate a target that appeared on the list from a lure. Thus, to detect the target word and avoid picking a semantically similar filler in a target-present word lineup, an adaptive strategy would be to discount the common semantic features shared by the words and to focus on more diagnostic semantic features (or perhaps even phonological features; Finley, Sungkhasettee, Roediger III, & Balota, 2017) that the target word would uniquely share with the memory trace. This theoretical interpretation offers one possible explanation for why, even when the stopping rule was removed from the sequential lineups, a simultaneous advantage was still evident (e.g., Fig. 5).

The inclusion of critical lures as attractive fillers in target-present and target-absent lineups (i.e., TPCL and TACL) had the effect of siphoning away some IDs that would have otherwise landed on a different word in the lineup. This can be seen in Tables 1 and 4 by comparing non-critical false-alarm rates for TP and TA to those for TPCL and TACL. Interestingly, the presence of a critical lure in the TP lineup did not have a detrimental effect on the hit rate with either type of lineup (compare hit rates for TP vs. TPCL); the critical lure only siphoned off responses from other fillers (for more on filler siphoning, see Smith, Wells, Smalarz, & Lampinen, 2018; Colloff, Wade, Strange, & Wixted, 2018).

The ROC curves based on hit rates from the TP lineups and critical false-alarm rates from the TACL lineups (Figs. 2C and 4C) provided an interesting scenario analogous to a biased lineup in the face memory domain: pitting lineups with a guilty suspect against lineups with an extremely suspicious

looking innocent suspect (the DRM critical lure). For this scenario, using pAUC as the measure of discriminability, the simultaneous test format yielded better performance than the sequential test format in Experiment 1 (but only without the stopping rule) and in Experiment 2.

Our overall finding is that a simultaneous test format yields better discriminability than a sequential test format. Although we have endeavored to provide every relevant statistical analysis, one really need only observe the ROC graphs (Figs. 2, 3, 4, and 5) to see that the simultaneous line is always above the sequential line. These laboratory experiments have extended lineup identification tasks to entirely different materials (word lists), with the added benefit of easily testing multiple lineups per subject. Traditional laboratory procedures and their accompanying analytical methods, particularly signal detection theory and ROC analyses, should continue to be employed to inform theory and data on important applied issues.

Open Practices Statement

The online Supplemental Materials include: data, R code used for analyses, word lists, instructions, and additional details about ROC construction and comparison. Neither experiment was pre-registered. Supplemental materials can be found online at either the publisher's website or the first author's website: jasonfinley.com

Author Note This research was supported by a James S. McDonnell Foundation 21st Century Science Initiative in Bridging Brain, Mind, and Behavior Collaborative Award to Henry L. Roediger, III. We are grateful to Laura Mickes and Scott D. Gronlund for helpful guidance in data analysis, and to Caren M. Rotello for helpful comments on earlier presentations of this work.

References

- 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.
- Buckhout, R. (1974). Eyewitness testimony. *Scientific American, 231*(6), 23–31.
- Carlson, C. A. & Carlson, M. A. (2014). An evaluation of perpetrator distinctiveness, weapon presence, and lineup presentation using ROC analysis. *Journal of Applied Research in Memory and Cognition, 3*, 45–53.
- Carlson, C. A., Carlson, M. A., Weatherford, D. R., Tucker, A., & Bednarz, J. (2016). The effect of backloading instructions on eyewitness identification from simultaneous and sequential lineups. *Applied Cognitive Psychology, 30*, 1005–1013. doi:<https://doi.org/10.1002/acp.3292>

- Clark, S. E. (2005). A re-examination of the effects of biased lineup instructions in eyewitness identification. *Law and Human Behavior, 29*(5), 575–604.
- Clark, S. E., & Godfrey, R. D. (2009). Eyewitness identification evidence and innocence risk. *Psychonomic Bulletin & Review, 16*(1), 22–42.
- Clifford, B. R., & Bull, R. (1978). *The psychology of person identification*. London: Routledge & Kegan Paul.
- Colloff, M. F., Wade, K. A., Strange, D., & Wixted, J. T. (2018). Filler-siphoning theory does not predict the effect of lineup fairness on the ability to discriminate innocent from guilty suspects: Reply to Smith, Wells, Smalarz, and Lampinen (2018). *Psychological Science, 29*(9), 1552–1557. doi:<https://doi.org/10.1177/0956797618786459>
- Deese, J. (1959). On the prediction of occurrence of particular verbal intrusions in immediate recall. *Journal of Experimental Psychology, 58*, 17–22. doi:<https://doi.org/10.1037/h0046671>
- 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.
- Finley, J. R., Roediger, H. I., Hughes, A. D., Wahlheim, C. N., & Jacoby, L. L. (2015). Simultaneous versus sequential presentation in testing recognition memory for faces. *The American Journal of Psychology, 128*(2), 173–195. doi:<https://doi.org/10.5406/amerjpsyc.128.2.0173>
- Finley, J. R., Sungkhasettee, V. W., Roediger III, H. L., & Balota, D. A. (2017). Relative contributions of semantic and phonological associates to over-additive false recall in hybrid DRM lists. *Journal of Memory and Language, 93*, 154–168.
- Police Executive Research Forum. (2013). *A national survey of eyewitness identification procedures in law enforcement agencies*. Retrieved from <http://policeforum.org/library/eyewitness-identification/NIJEyewitnessReport.pdf>
- Garcia, M. A. (n.d.). *Collector: A program for running psychology experiments on the web*. Retrieved from <https://github.com/gikeymarcia/Collector>
- Garrett, B. F. (2011). *Convicting the innocent: Where criminal prosecutions go wrong*. Cambridge, MA: Harvard University Press.
- 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.
- 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:<https://doi.org/10.1177/0963721413498891>
- Gross, S. R., & Shaffer, M. (2012). *Exonerations in the United States, 1989–2012: Report by the National Registry of Exonerations*. Retrieved from https://www.law.umich.edu/special/exoneration/Documents/exonerations_us_1989_2012_full_report.pdf
- Horry, R., Palmer, M. A., & Brewer, N. (2012). Backloading in the sequential lineup prevents within-lineup criterion shifts that undermine eyewitness identification performance. *Journal of Experimental Psychology: Applied, 18*, 346–360. doi:<https://doi.org/10.1037/a0029779>
- Lindsay, R. C., & Wells, G. L. (1985). Improving eyewitness identifications from lineups: Simultaneous versus sequential lineup presentation. *Journal of Applied Psychology, 70*, 556–564. doi:<https://doi.org/10.1037/0021-9010.70.3.556>
- Lindsay, R. C. L., Lea, J. A., Nosworthy, G. J., Fulford, J. A., Hector, J., LeVan, V., & Seabrook, C. (1991). Biased lineups: Sequential presentation reduces the problem. *Journal of Applied Psychology, 76*(6), 796.
- Macmillan, N. A., & Creelman, C. D. (2004). *Detection theory: A user's guide* (2nd ed.). doi:<https://doi.org/10.4324/9781410611147>
- Mansour, J. K., Beaudry, J. L., & Lindsay, R. C. L. (2017). Are multiple-trial experiments appropriate for eyewitness identification studies? Accuracy, choosing, and confidence across trials? *Behavior Research Methods, 49*, 2235–2254. <https://doi.org/10.3758/s13428-017-0855-0>
- Meissner, C. A., Tredoux, C. G., Parker, J. F., & MacLin, O. H. (2005). Eyewitness decisions in simultaneous and sequential lineups: A dual-process signal detection theory analysis. *Memory & Cognition, 33*(5), 783–792.
- Meisters, J., Diederhoben, B., & Musch, J. (2018). Eyewitness identification in simultaneous and sequential lineups: An investigation of position effects using receiver operating characteristics. *Memory*, DOI: <https://doi.org/10.1080/09658211.2018.1464581>
- Mickes, L., Flowe, H. D., & Wixted, J. T. (2012). Receiver operating characteristic analysis of eyewitness memory: Comparing the diagnostic accuracy of simultaneous versus sequential lineups. *Journal of Experimental Psychology: Applied, 18*, 361–376. doi:<https://doi.org/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*(2), 58–62.
- Munsterberg, H. (1908). *On the witness stand: Essays on psychology and crime*. doi:<https://doi.org/10.1037/10854-000>
- Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). The University of South Florida word association, rhyme, and word fragment norms. Retrieved from <http://www.usf.edu/FreeAssociation/>
- Palmer, M. A., & Brewer, N. (2012). Sequential lineup presentation promotes less-biased criterion setting but does not improve discriminability. *Law and Human Behavior, 36*(3), 247.
- Pardilla-Delgado, E., & Payne, J. D. (2017). The Deese-Roediger-McDermott (DRM) task: A simple cognitive paradigm to investigate false memories in the laboratory. *Journal of Visualized Experiments, 119*, 54793. doi:<https://doi.org/10.3791/54793>
- Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J. C., & Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics, 12*(1), 77. doi:<https://doi.org/10.1186/1471-2105-12-77>
- Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*, 803–814. doi:<https://doi.org/10.1037/0278-7393.21.4.803>
- Roediger, H. L., Watson, J. M., McDermott, K. B., & Gallo, D. A. (2001). Factors that determine false recall: A multiple regression analysis. *Psychonomic Bulletin & Review, 8*, 385–407. doi:<https://doi.org/10.3758/BF03196177>
- Rotello, C. M., & Chen, T. (2016). ROC curve analyses of eyewitness identification decisions: An analysis of the recent debate. *Cognitive Research: Principles and Implications*. DOI <https://doi.org/10.1186/s41235-016-0006-7>.
- Rotello, C. M., & Macmillan, N. A. (2007). Response bias in recognition memory. *Psychology of Learning and Motivation, 48*, 61–94.
- Smith, A. M., Wells, G. L., Smalarz, L., & Lampinen, J. M. (2018). Increasing the similarity of lineup fillers to the suspect improves the applied value of lineups without improving memory performance: Commentary on Colloff, Wade, and Strange (2016). *Psychological Science, 29*(9), 1548–1551.
- Stebly, 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:<https://doi.org/10.1037/a0021650>
- Sučić, I., Tokić, D., & Ivešić, M. (2015). Field study of response accuracy and decision confidence with regard to lineup composition and lineup presentation. *Psychology, Crime & Law, 21*, 798–819. doi:<https://doi.org/10.1080/1068316X.2015.1054383>
- Terrell, J. T., Baggett, A. R., Dasse, M. N., & Malavanti, K. F. (2017). Hybridization of simultaneous and sequential lineups reveals

- diagnostic features of both traditional procedures. *Applied Psychology in Criminal Justice*, 13, 97–109.
- Wells, G. L. (1984). The psychology of lineup identifications. *Journal of Applied Social Psychology*, 14, 89–103. doi:<https://doi.org/10.1111/j.1559-1816.1984.tb02223.x>
- Wells, G. L. (2001). Eyewitness lineups: data, theory, and policy. *Psychology, Public Policy, and Law*, 7, 791–801. doi:<https://doi.org/10.1037/1076-8971.7.4.791>
- Wells, G. L., Smalarz, L., & Smith, A. M. (2015). ROC analysis of lineups does not measure underlying discriminability and has limited value. *Journal of Applied Research in Memory and Cognition*, 4(4), 313–317. doi:<https://doi.org/10.1016/j.jarmac.2015.08.008>
- Wells, G. L., Small, M., Penrod, S., Malpass, R. S., Fulero, S. M., & Brimacombe, C. E. (1998). Eyewitness identification procedures: Recommendations for lineups and photospreads. *Law and Human Behavior*, 22(6), 603.
- Wells, G. L., Smith, A. M., & Smalarz, L. (2015). ROC analysis of lineups obscures information that is critical for both theoretical understanding and applied purposes. *Journal of Applied Research in Memory and Cognition*, 4(4), 324–328. doi:<https://doi.org/10.1016/j.jarmac.2015.08.010>
- Wilson, B. M., Donnelly, K., Christenfeld, N., & Wixted, J. T. (2019). Making sense of sequential lineups: An experimental and theoretical analysis of position effects. *Journal of Memory and Language*, 104, 108–125.
- Wixted, J. T., & Mickes, L. (2012). The field of eyewitness memory should abandon probative value and embrace receiver operating characteristic analysis. *Perspectives on Psychological Science*, 7(3), 275–78.
- Wixted, J. T., & Mickes, L. (2014). A signal-detection-based diagnostic-feature-detection model of eyewitness identification. *Psychological Review*, 121(2), 262.
- Wixted, J. T., & Mickes, L. (2015). Evaluating eyewitness identification procedures: ROC analysis and its misconceptions. *Journal of Applied Research in Memory and Cognition*, 4(4), 318–323.
- Wixted, J. T., Vul, E., Mickes, L. & Wilson, B. W. (2018). Models of lineup memory. *Cognitive Psychology*, 105, 81–114.
- Zhu, B., Chen, C., Loftus, E. F., Lin, C., & Dong, Q. (2013). The relationship between DRM and misinformation false memories. *Memory & Cognition*, 41(6), 832–838.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.