

Beyond the Confidence-Accuracy Relation: A Multiple-Reflector-Variable Approach to Postdicting Accuracy on Eyewitness Lineups

Nydia T. Ayala¹, Andrew M. Smith², and Gary L. Wells²

¹ Department of Cognitive and Behavioral Science, Washington and Lee University

² Department of Psychology, Iowa State University

We examined whether the potential to distinguish between accurate and inaccurate decisions on eyewitness lineups could be improved by combining information from three witness behaviors: confidence, decision time, and the language that witnesses use to justify their lineup decisions. We assessed the postdictive potential of these variables for both positive identifications and lineup rejections on both simultaneous and sequential lineups. All three behaviors independently postdicted the accuracy of both positive identifications and lineup rejections for both simultaneous and sequential lineups. The potential to distinguish between accurate and inaccurate lineup decisions was maximized by considering all three variables. Interestingly, the classifier trained to distinguish the language of accurate and inaccurate witnesses appeared to recover a distinction between use of absolute- and relative-judgment strategies. For both simultaneous and sequential lineups, accurate decisions were accompanied by absolute language and inaccurate decisions were accompanied by relative language. The applied implications of this work are clear—accurate witnesses are confident, fast, and reference an absolute-judgment strategy. This work also advances theory on why sequential lineups lead to worse discriminability than do simultaneous lineups. Sequential lineups do not increase use of absolute-judgment strategies, but might make it more difficult to determine the strongest match to memory.

Public Significance Statement

Outside observers trusted with evaluating the accuracy of eyewitness identification decisions would do well to consider not only eyewitness confidence, but also decision time, and the language that witnesses use to justify their identification decisions. Although confidence is strongly associated with lineup-identification accuracy, it does not perfectly distinguish accurate from inaccurate witnesses, and by considering decision time, as well as the language that witnesses use to justify their identification decisions, outside observers would improve their ability to distinguish accurate from inaccurate identification decisions. Accurate witnesses are confident, fast, and tend to use absolute language to justify their identification decisions. In contrast, inaccurate witnesses are unconfident, slow, and tend to use relative language to justify their identification decisions.

Keywords: eyewitness memory, simultaneous lineups, sequential lineups, confidence-accuracy relation, reflector variables

Supplemental materials: <https://doi.org/10.1037/xap0000527.supp>

This article was published Online First March 3, 2025.

Deanna Kuhn served as action editor.

Nydia T. Ayala  <https://orcid.org/0000-0003-4073-8141>

Andrew M. Smith  <https://orcid.org/0000-0002-4184-9364>

The data and analytic code for these experiments are available on the Open Science Framework at https://osf.io/x2ru4/?view_only=bcc3d6de9cec4df380632d4ff4bc21ec. The authors have no conflicts of interest to disclose. This work was supported by grants awarded to Andrew M. Smith and Gary L. Wells from the National Science Foundation (Grant 2017510) and from the Laura and John Arnold Foundation (Grant 400-61-12C). The views and opinions expressed are those of the authors and do not necessarily reflect the views of the funders. The authors thank Ian Dobbins and Justin Kantner for making their extremely useful R scripts openly available. The

authors also thank Zohara Assadipour for her help in preprocessing witness-participants' justification statements.

Nydia T. Ayala played a lead role in data curation, formal analysis, investigation, project administration, software, visualization, and writing—original draft and an equal role in methodology. Andrew M. Smith played a lead role in conceptualization, funding acquisition, and supervision, a supporting role in methodology, and an equal role in writing—review and editing. Gary L. Wells played a lead role in conceptualization and funding acquisition, a supporting role in writing—review and editing, and an equal role in methodology.

Correspondence concerning this article should be addressed to Nydia T. Ayala, Department of Cognitive and Behavioral Science, Washington and Lee University, Parnly 234, 204 West Washington Street, Lexington, VA 24450, United States. Email: nayala@wlu.edu

During criminal investigations law enforcement personnel often present witnesses with lineups. A lineup is a procedure in which a suspect is surrounded by known-innocent persons (*fillers*) and presented to a witness for an identification attempt. Critically, suspects are often innocent, and police investigators do not know whether a given suspect is innocent or guilty. The purpose of a lineup is to gather information from the witness that can shed light on the likely guilt or innocence of the suspect (A. M. Smith et al., 2020; Wells & Luus, 1990). Whereas an identification of the suspect provides evidence of guilt, a rejection of the lineup provides evidence of innocence (Ayala et al., 2022; A. M. Smith et al., 2023; Wells, Yang, & Smalarz, 2015). Yet, human memory is fallible; sometimes witnesses mistakenly identify innocent suspects and other times witnesses fail to identify guilty suspects. There are two approaches to combatting these errors. The first involves developing identification procedures that maximize guilty-suspect identifications while minimizing innocent-suspect identifications (Wells et al., 1998, 2020). This *prospective* approach is intended to prevent mistakes before they happen. But it is impossible to eliminate mistakes entirely, and so the legal system also needs a *retrospective* approach for distinguishing between likely accurate and likely inaccurate decisions. In the present work, we examine the potential for *reflector variables* to retrospectively distinguish between accurate and inaccurate lineup decisions (e.g., Quigley-McBride & Wells, 2023; A. M. Smith & Wells, 2023; Wells, 2020).

A reflector variable is any behavior demonstrated by a witness during an identification procedure that covaries with the guilt or innocence of the suspect (Wells, 2020). The reflector-variable approach to estimating witness accuracy is a straightforward extension of the logic for showing a lineup to a witness. The logic for showing a lineup to a witness is that a witness' decision is mediated by the strength of match between the police suspect and her memory for the culprit. Because a guilty suspect will typically provide a stronger match to the witness' memory for the culprit than will an innocent suspect, a witness' decision should depend on, or *reflect*, the guilt or innocence of the suspect. But categorical lineup decisions provide only a crude reflection of guilt or innocence (N. Brewer et al., 2012; Sauer, Brewer, & Weber, 2008). There are many other behaviors that a witness might display during a lineup that provide nuanced information about the likely guilt or innocence of the suspect. For example, decision time, vocalized decision processes, and expressed level of confidence all reflect the likely accuracy of a witness' identification decision. Whereas innocent-suspect identifications are associated with references to deliberative and comparative processes (Behrman & Richards, 2005; Dunning & Stern, 1994), guilty-suspect identifications are made more quickly and with greater confidence (e.g., N. Brewer & Wells, 2006; Dunning & Perretta, 2002; Quigley-McBride & Wells, 2023; Sporer et al., 1995; Weber et al., 2004; Wixted & Wells, 2017).

The most studied reflector variable by far is eyewitness confidence. This is with good reason. Confidence judgments are easy to collect, and people have experience rendering confidence judgments across a variety of domains ranging from decision-making (e.g., Ackerman & Thompson, 2014) to perceptual judgments (e.g., Shekhar & Rahnev, 2024; Stephens et al., 2017) and memory judgments (e.g., Bjork et al., 2013; Dunlosky et al., 2005; Metcalfe & Finn, 2008; Wixted & Wells, 2017). Most importantly, across these domains, confidence tends to be highly correlated with accuracy

(W. F. Brewer et al., 2005; DeSoto & Roediger, 2014; Dunlosky & Metcalfe, 2009; Koriart & Goldsmith, 1996; Perfect & Hollins, 1996; Thompson et al., 2011). In the context of eyewitness lineups, confidence does an excellent job of sorting between guilty-suspect identifications and innocent-suspect identifications (Palmer et al., 2013; Sporer et al., 1995; Wixted & Wells, 2017). Indeed, the current authority on eyewitness identification concluded that when lineups adhere to science-based best practice recommendations, high confidence implies high accuracy and low confidence implies low accuracy (Wixted & Wells, 2017).

Although confidence might imply accuracy, it certainly does not guarantee it. In fact, some recent work estimates high-confidence accuracy to be lower than what was observed by Wixted and Wells (2017; Fitzgerald et al., 2023), especially when viewing conditions are poor (Giacona et al., 2021). Furthermore, only a fraction of witnesses who encounter culprit-present lineups identify the culprit with high confidence (e.g., Palmer et al., 2013; Quigley-McBride & Wells, 2023; Wells, Yang, & Smalarz, 2015). In other words, for most witnesses (viz. those who do not identify the suspect with high confidence), confidence is insufficient for determining the likely guilt of the suspect. We bring up these limitations to eyewitness expressions of confidence not to discount the value of confidence statements, but to suggest that something more is needed. The criminal justice system arguably stands to gain from a multiple-reflector-variable approach to postdicting accuracy. There is ample room for other reflector variables beyond confidence to work additively or multiplicatively with confidence to improve our ability to postdict the accuracy of lineup decisions. Consistent with this notion, the handful of studies that have looked at the utility of using multiple reflector variables in concert, have typically found improved postdiction compared to using confidence in isolation (e.g., N. Brewer & Wells, 2006; Dodson & Dobolyi, 2016; Grabman et al., 2019; Price et al., 2020).

Beyond confidence, decision time is the second most studied reflector variable. The shorter the decision time, the more likely it is that the witness made an accurate decision (N. Brewer et al., 2006; Dunning & Perretta, 2002; Sauer, Brewer, & Wells, 2008; Sauerland & Sporer, 2007, 2009; S. M. Smith et al., 2000; Sporer, 1992, 1993; Weber et al., 2004). By no means are confidence and decision time the only two reflector variables of value. There are also marked differences in the self-reported decision strategies of accurate and inaccurate witnesses. For example, consider the findings of Dunning and Stern (1994). After running witness-participants through a lineup procedure, the experimenters asked witness-participants to select from a list of eight statements, those that best reflected their decision-making process during the lineup. Whereas accurate witnesses endorsed statements that implied a relatively automatic decision-making process (e.g., "His face just 'popped out' at me."), inaccurate witnesses endorsed statements that implied a more deliberative decision-making process (e.g., "I first eliminated the ones that definitely were not him, then choose among the rest."). More recently, Mansour et al. (2024) employed a more extensive, 17-item questionnaire on witness decision-making process and found that self-reported levels of automaticity and self-reported judgment strategies both correlated with suspect-identification accuracy.

But asking witnesses to self-report their decision processes by responding to a series of multiple-choice questions is artificial and likely constrains what can be learned about the witness' decision. Recent innovations have led to asking witnesses to justify their

decision or expression of confidence in an open-ended statement (Dobbins, 2022; Dobbins & Kantner, 2019; Seale-Carlisle et al., 2022). These statements are then passed on to machine learning algorithms for a finer grained analysis on the language that characterizes accurate and inaccurate decisions. For example, participants in Dobbins and Kantner (2019) completed an old–new word recognition task and on a subset of trials, were asked to justify their accompanying “old” or “new” decision. The authors applied machine learning analyses to the justification statements and found that the language accompanying hits largely reflected a recollective experience, whereas the language accompanying false alarms largely reflected the absence of a recollective experience. What is more, the language used in these justifications better discriminated between hits and false alarms than did decision time. In the context of eyewitness identification research, words that convey high confidence are associated with hits, and words that convey low confidence are associated with false alarms (Seale-Carlisle et al., 2022). Critically, the language used in these confidence justifications provided uniquely predictive information beyond that of numeric confidence.

The Asymmetric Utility of Reflector Variables

Although confidence has proven useful for distinguishing between guilty-suspect identifications and innocent-suspect identifications, it has proven far less useful for distinguishing between correct rejections and incorrect rejections (e.g., N. Brewer & Wells, 2006; Sporer et al., 1995; Wixted & Wells, 2017). Two general explanations have been put forth to explain why confidence is more strongly related to the accuracy of suspect identifications than to the accuracy of rejections. Some have suggested that it is simply more difficult to detect the absence of something than it is to detect the presence of that same thing (e.g., Clark & Wells, 2008; Wells, Yang, & Smalarz, 2015). Others have suggested that there are differences in how choosers and rejecters generate confidence judgments and that the process used by rejecters is suboptimal (e.g., R. C. L. Lindsay et al., 2013; Sauerland et al., 2012; Yilmaz et al., 2022). According to this dual-process account, suspect-identification confidence is based on the degree of match between the suspect and the witness’ memory for the culprit, but rejection confidence is based on averaging the match strength of all lineup members to form a global confidence rating. This averaging process is assumed to be suboptimal in some way that attenuates the strength of the confidence-accuracy relation. We revisit the discussion of what causes this asymmetric utility in the Discussion section and offer yet another alternative explanation.

Our primary concern in the present article lies not in explaining why reflector variables are more useful for affirmative identification decisions than for rejection decisions, but in determining whether there might be some conditions in which reflector variables bear a strong relation to the accuracy of rejection decisions. There is some research demonstrating that confidence better distinguishes between culprit identifications and innocent-suspect identifications when there is variability across witnesses in encoding conditions compared to when encoding conditions are held constant (D. S. Lindsay et al., 1998, 2000; Molinaro et al., 2021). The logic is that when encoding conditions are held constant, all witnesses have roughly the same degree of memory strength, roughly the same probability of making correct decisions, and that this lack of between-witness variability leads witnesses to render their identification decisions with similar

levels of confidence. This same logic is equally applicable to rejection decisions. By varying the quality of encoding conditions, we can create witnesses with strong memories and witnesses with weak memories. When a witness with a strong memory has a weak recognition experience, that is likely because the culprit is absent from the lineup, and the witness should be relatively likely to attribute the weak recognition experience to culprit absence. But when a witness with a weak memory has the same weak recognition experience, it is less clear whether that is attributable to culprit absence or to weak memory, and the witness should be uncertain about the source of the weak recognition experience. Hence, variations in memory strength should lead to both variations in accuracy and behavior.

We manipulated the quality of encoding conditions (via simulated viewing distance) in the present experiment for two reasons. First, consistent with past work, reflector variables might better postdict the accuracy of rejection decisions in the context of variable encoding conditions. Second, as we discuss next, we also examined the language that witnesses used to justify their lineup decisions and the potential for machine learning models to distinguish between the language of accurate and inaccurate witnesses. Variability in the quality of viewing conditions should lead to variability in the language that witnesses use to justify their lineup decisions, which is important both for saturating the range of justifications that witnesses might provide and for generalizations to actual criminal investigations, where memory varies drastically from one witness to the next (Perfect, 2002).

Examining the Language That Witnesses Use to Justify Their Lineup Decisions

The language that witnesses use to justify their lineup decisions has great potential to shed light on both applied and theoretical questions related to eyewitness identification. To maximize the potential applied implications of decision justifications, it is important that experiments capture the essential components of real-world identification procedures. That memory strength varies considerably from one witness to the next likely has tremendous impact on the language that witnesses use to justify their lineup decisions in actual cases. In theory, a witness with a strong memory should find the lineup task easy, which might be reflected by a confident, vivid, and resolute justification. A witness with a weak memory is just the opposite. That witness should find the lineup task difficult, which might be reflected by an uncertain, equivocal, and deliberative justification. Of course, witnesses with strong and weak memories differ not only in their justifications but also in the likelihood of making accurate lineup decisions. Whereas witnesses with strong memories are relatively likely to make accurate lineup decisions, witnesses with weak memories are relatively likely to make inaccurate lineup decisions. Because we wanted to capture as much as possible, the variability in the language witnesses use to justify their lineup decisions, we manipulated memory strength via the quality of encoding conditions.

From a theoretical perspective, there are many potential insights that might be gleaned from examining the language that witnesses use to justify their lineup decisions. For example, the language that individuals use to justify their identification decisions distinguishes between the presence and absence of recollective experiences (Dobbins & Kantner, 2019). It stands to reason that the language

witnesses use to justify their lineup decisions might similarly distinguish between absolute- and relative-judgment strategies. The distinction between absolute- and relative-judgment strategies has long driven theorizing in the context of eyewitness identification research (e.g., Wells, 1984, 1993). The relative-judgment strategy conceptualization was introduced to explain why witnesses were relatively likely to mistakenly identify innocent persons who bore only a modest resemblance to the culprit. At the extreme, a witness who adopts a relative-judgment strategy determines which lineup member provides the strongest match to memory and identifies that person without considering the absolute match between that person and her memory for the culprit. Arguably, what the witness should do instead is adopt an absolute-judgment strategy and base her decision on the absolute strength of match between the best matching lineup member and her memory for the culprit. In signal detection terms, a relative-judgment strategy can be thought of as adopting an overly lax decision criterion.

The distinction between absolute- and relative-judgment strategies also led to the development of the *sequential lineup procedure* (R. C. L. Lindsay & Wells, 1985; R. C. L. Lindsay et al., 1991; Steblay et al., 2011). Up until that point, the default method for conducting lineups was to present all lineup members to the witness simultaneously—or all at once. But simultaneous lineups make it easy for witnesses to compare faces and to determine which lineup member provides the strongest match to memory, which in theory, could promote the adoption of an overly lax decision criterion (i.e., relative-judgment strategy). Instead, R. C. L. Lindsay and Wells (1985) proposed that lineups should be conducted sequentially. With the sequential lineup procedure, the lineup members are presented to the witness one at a time, the witness must decide about member *i* before proceeding to member *i* + 1, the witness is not permitted to return to a lineup member that she has already decided on, and the witness should not know how many members are included in the lineup. In theory, the sequential procedure should encourage the witness to base her decision on the absolute match between a lineup member and her memory for the culprit and should discourage relative-judgment strategies. But do they? Examining the language that witnesses use to justify their decisions could provide insight on the efficacy of sequential lineups at promoting absolute-judgment strategies. To the extent that sequential lineups promote absolute-judgment strategies and simultaneous lineups promote relative-judgment strategies, we would expect to see more absolute language in the justifications for sequential lineups and more relative language in the justifications for simultaneous lineups. Moreover, because both simultaneous and sequential procedures are used in actual criminal investigations, this experiment offers an important test of whether the language witnesses use to justify their decisions generalizes across these two procedures.

The Present Study

The purpose of this work is to move beyond evaluating the utility of reflector variables in isolation and instead to assess the potential for combinations of reflector variables to sort accurate identifications from inaccurate identifications and accurate rejections from inaccurate rejections. Hence, we measured decision time, collected postidentification confidence statements, and importantly, asked participants to justify their lineup decision in their own words. We then subjected these statements to machine learning analyses. To maximize

generalizability to actual cases, we created variability in memory strength for the culprit by incorporating multiple simulated viewing distances at encoding. Moreover, because both simultaneous and sequential lineups are used in practice (Police Executive Research Forum, 2013), we included both presentation formats in this experiment.

Method

Transparency and Openness

We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study. We preregistered this experiment and the data analysis plan prior to data collection and analysis. This preregistration (https://osf.io/qwzkb?view_only=8102abb827ec4312b6ed6ddc2f7cc559) as well as de-identified data and analytic code are available on the Open Science Framework (https://osf.io/x2ru4/?view_only=bcc3d6de9cec4df380632d4ff4bc21ec; Ayala et al., 2021). The institutional review board at Iowa State University approved this experiment.

Participants

A total of 7,158 Mechanical Turk workers participated in this study in exchange for \$0.50 USD. We required that participants be 18 years or older, reside in the United States of America, and be fluent in English to participate. To better ensure quality of responses, we required that workers have a Human Intelligence Task approval rating of 80% or higher and have completed at least 50 Human Intelligence Tasks in order to participate. We used the TurkPrime platform to collect data (Litman et al., 2017). As noted in our preregistration, we intended to collect data from 6,000 participants,¹ but recruited an additional 1,039 participants in response to a programming error that we detected after reaching the preregistered sample size.² We proceeded with data analyses only after completing participant recruitment entirely.

Of the 7,158³ Mechanical Turk participants who completed this study, we excluded 928 participants, because they failed a bot check (*n* = 26); failed or did not answer the attention check (*n* = 304); encountered a technological problem with the encoding video or lineup (*n* = 218); recognized the stimuli from another experiment (*n* = 7); failed to submit a completion code for remuneration (*n* = 20); completed the study more than once (*n* = 9); submitted an unauthorized completion code (*n* = 3); completed the study using an

¹ Our primary research interest involves using machine learning to evaluate the language of lineup justifications. Unlike most analytic approaches in psychological science, machine learning analyses do not involve testing a priori hypotheses. Accordingly, power analyses do not exist for machine learning analyses. Our sample size was determined as the largest sample size we could feasibly afford to collect that best accommodated our moderately sized design.

² The programming error was restricted to the culprit-present sequential lineup condition. The error caused the culprit to appear only in the first or last position of the sequential lineup and never in the second, third, fourth, or fifth positions. Accordingly, we collected data from an additional 1,000 participants and assigned each of them to view a culprit-present sequential lineup where the culprit was always in one of positions two, three, four, or five.

³ Due to a subset of MTurk workers who completed the experiment without fully completing a request for remuneration and a separate subset who received compensation without completing the experiment, the recruited sample size deviates slightly from the 7,000 participation slots we allocated for.

unauthorized browser ($n = 27$); provided responses reflecting careless or insufficient effort (such as responding to open-ended questions with “great” or “nice,” or providing unintelligible responses; $n = 211$); failed to provide an identification decision ($n = 1$); or identified three or more individuals from the sequential lineup procedure ($n = 102$), leaving us with a final sample of 6,230 participants. Among those participants included in the final sample, 2,773 were randomly assigned to the simultaneous lineup condition, and 3,457 were randomly assigned to the sequential lineup condition.

Of those participants who reported demographic information, the reported ages ranged from 18 to 90 years ($M = 37.76$, $SD = 12.35$); 61.96% were female, 37.32% were male, and 0.72% of participants reported “other” as their gender. With regard to participants’ racial or ethnic background, the following breakdown was observed: White or Caucasian (72%), Black or African American (9.29%), Asian or Asian American (8.10%), Hispanic or Latino/Latina (7.26%), other (2.01%), Middle Eastern (0.55%), Native American (0.55%), and Native Hawaiian or other Pacific Islander (0.26%).

Design

We randomly assigned participants to a 5 (encoding distance: 10, 60, 110, 160, 210 ft.) $\times$ 2 (lineup presentation: simultaneous, sequential) $\times$ 2 (culprit: present, absent) between-participants design. For purposes of stimulus sampling (Wells & Windschitl, 1999), participants were randomly assigned to view one of three target actors (from one of five simulated viewing distances). After viewing the encoding video, participants completed a culprit-present or culprit-absent simultaneous or sequential lineup procedure. Participants qualified their lineup decision with an expression of confidence, which they provided on an 11-point scale that ranged from 0% to 100%. After providing their expression of confidence, participants were asked to explain in their own words why they made the decision they made. The entire experiment was programmed and completed in Qualtrics.

Materials

Crime Videos

Participants viewed a short video depicting one of three target actors speaking to someone off camera. The videos did not have any sound, and the camera frame captured the target actors from the chest up. The videos were 16 s in length, including a 3-s-long blank screen at both the start and end of the video, which allowed time for the video to buffer prior to the presentation of the target. We postedited existing videos of each target actor to reflect how the target would have been viewed from various distances. The purpose of this manipulation was not to examine the impact viewing distance has on lineup performance, but rather to increase the variance in participants’ memory strength for the culprit. Manipulating viewing distance accomplishes this because as the distance between an eyewitness and the target’s face increases, the perceptual information that is required to form a strong memory trace for the target’s face becomes less available, impairing recognition performance in turn (Davis & Peterson, 2022; Lampinen et al., 2014; R. C. L. Lindsay et al., 2008; Loftus & Harley, 2005; Wagenaar & van der Schrier, 1996). Details on the video postediting process are available in Supplemental Materials.

We pilot tested the videos to determine the shortest and longest viewing distance that we should include in the primary experiment. We determined that the shortest distance would be the simulated viewing distance that resulted in a high rate of culprit identification decisions (e.g., 80%), and that the longest distance would be the simulated viewing distance that resulted in a near-chance culprit-identification rate. To add additional variation, we included multiple other distances between the shortest and longest distance that produced different culprit-identification rates. Based on pilot testing, we modified the encoding videos to approximate the following five distances: 10, 60, 110, 160, and 210 ft. Hence, the main experiment included 15 videos (five videos for each of the three target individuals). The longest viewing distance is in line with research showing that eyewitness identification performance drops to chance levels somewhere between 150 and 200 ft. (Lampinen et al., 2014; Nyman et al., 2019).

Eyewitness Lineups

Participants viewed a culprit-present or culprit-absent six-person lineup that was presented in either a simultaneous or sequential format. For culprit-present lineups, the culprit was presented alongside five novel filler individuals; while the culprit-absent lineups were comprised of six novel filler individuals (we did not use a designated-innocent suspect). Each target was designated 12 different fillers that fit the general description of that target. Both simultaneous and sequential lineup conditions used the same filler and culprit photographs. For culprit-present lineups, Qualtrics randomly selected five of the 12 fillers to appear alongside the photo of the target, and for culprit-absent lineups, Qualtrics randomly selected six of the 12 fillers to appear in the lineup. For all lineup conditions, Qualtrics also randomized the position of all individuals in the lineup. The selection of fillers and randomization of lineup member positions was done anew for each participant.

Procedure

After providing informed consent, we instructed participants that they would watch a video and because they would only have one opportunity to watch the video, they should only advance the page when ready. We also instructed participants that the video would not have any audio and that they should play the video in full screen.

After the video finished playing, participants completed a 2-min-long anagram filler task and were then presented with instructions for the lineup task. We admonished participants that the culprit may or may not be present in the lineup, and we further advised participants that their selection from the lineup would automatically advance the screen. We programmed this function into the Qualtrics survey to increase the accuracy of the decision time measure. Both the sequential and simultaneous lineup procedures were self-paced. The lineup instructions provided to participants are detailed below.

Participants assigned to the simultaneous lineup procedure read the following instructions prior to viewing the lineup:

You saw a person in the video that you viewed earlier. You are now going to view a set of photos. All of the photos will be presented at the same time. If the person you viewed in the video is present in the lineup, please select his photo, otherwise, please select “Not Present.” Keep in mind that the person from the video may or may not be present in the set

of photos that you are about to view. Note: The lineup page will advance automatically as soon as you click an option.

On the page in which the lineup appeared, participants were instructed with: “If the person who you saw in the video is present, select his photo, otherwise, select ‘Not Present.’”

Participants assigned to the sequential lineup procedure read the following instructions prior to viewing the lineup:

You saw a person in the video that you viewed earlier. You are now going to view a set of photos. The photos will be presented one at a time and in a random order. For each photo you will be asked whether the person appearing in the photo is the man you saw in the video. You are going to be shown all of the photos regardless of your decision. You must make a decision about each photo before moving on to the next one. Keep in mind that the person from the video may or may not be present in the set of photos that you are about to view. Note: The lineup pages will advance automatically as soon as you click an option.

On the pages in which sequentially presented lineup members appeared, participants were instructed with: “If this is the person that you saw in the video, please select his photo, otherwise, please select ‘Not Him.’”

After making a lineup decision, witnesses were asked to express their level of confidence in that decision on an 11-point scale ranging from 0% to 100%. For the simultaneous lineup, this confidence rating was always on the page immediately following the lineup procedure. For the sequential lineup, if a witness identified someone, they were asked to indicate their confidence before viewing any additional lineup members. If a witness rejected all lineup members, they were asked to indicate their confidence that the culprit was not present in the lineup after making their final rejection. It is possible for witnesses to identify multiple persons from a sequential lineup procedure. We remedied multiple affirmative responses by first excluding those who identified three or more individuals, and for those witnesses who identified two individuals, taking the identification that was made with higher confidence. In the case that a witness made two identifications with the *same* confidence, we took the last of those identifications.

After providing their confidence rating, all participants were then asked to explain in their own words, why they made this decision (“In your own words, please explain why you made this decision.”). Participants provided their response to this prompt via a text box entry. Every participant in the simultaneous lineup condition as well as those participants who rejected the entire sequential lineup, provided an explanation for their decision immediately following their confidence rating. Participants who made an affirmative identification from a sequential lineup provided their justification after viewing any remaining lineup members (if applicable).⁴ Upon completion of the experimental session, participants provided demographic information and answered an attention check question. Afterward, participants specified whether they experienced any technical difficulties and were then fully debriefed.

Results

We begin this section by outlining descriptive statistics for both simultaneous and sequential lineup conditions. Following this, we detail how we constructed the language classifier and examine which language proved most useful for distinguishing between correct and incorrect lineup decisions. Finally, we examine whether using

multiple reflector variables in concert improve our ability to distinguish between correct and incorrect lineup decisions.

Descriptive Statistics

Table 1 displays the proportion of suspect identifications, filler identifications, and rejection decisions as a function of lineup presentation format, distance, and culprit presence condition. In addition, Table 1 also displays the discriminability index (d' ; Green & Swets, 1966; Swets, 1973) at each simulated viewing distance for both simultaneous and sequential lineups. Although not the primary purpose of the present work, it is notable that at every single distance, the simultaneous lineup had better discriminability than did the sequential lineup (Dobolyi & Dodson, 2013; Mickes et al., 2012). Fitting with decades of research comparing simultaneous and sequential lineups, the simultaneous lineup led to a higher hit rate than the sequential lineup at every single distance (e.g., Steblay et al., 2001, 2011). What is perhaps more surprising is that the simultaneous lineup also led to, if anything, more correct rejections and fewer innocent-suspect identifications than did sequential lineups. Given that simultaneous lineups led to both more culprit identifications and more correct rejections than did sequential lineups, there is no debate that simultaneous lineups have better discriminability.

A concern with sequential lineups is the possibility of position effects. A position effect occurs when witness discriminability differs systematically across the various lineup positions (Gronlund et al., 2012; Wells, Steblay, & Dysart, 2015). Table 2 shows a breakdown of culprit identification rates, culprit-absent false-alarm rates, and d' across the various positions of the sequential lineup. Given that there is no evidence that discriminability differs systematically across positions in the sequential lineup, we do not consider lineup position further in our analyses.

Building the Language Classifiers

Consistent with past research, we used a “bag-of-words” approach to build our language classifiers (Dobbins & Kantner, 2019; Seale-Carlisle et al., 2022). Each word used by at least one witness was treated as a potential predictor variable that could be used to distinguish between accurate and inaccurate responses. We carried out the analyses we report below in R (v4.4.1; R Core Team, 2024) and RStudio (RStudio Team, 2020) with various packages: tidyverse (Wickham et al., 2019), dplyr (Wickham et al., 2023), glmnet (Friedman et al., 2010), quanteda (Benoit et al., 2018), tm (Feinerer et al., 2008; Feinerer & Hornik, 2023), caret (Kuhn, 2008), gridExtra (Baptiste, 2015), readtext (Benoit & Obeng, 2021), ggpubr (Kassambara, 2023), and pROC (Robin et al., 2011).

Prior to building the classifiers, we preprocessed the justification statements in preparation for subsequent analyses. We first removed any punctuation and resolved any obvious grammatical or spelling errors. No other edits were made to the words. After this, we split the data into separate models based on the witness’ decision and lineup presentation format (simultaneous or sequential). In an actual police lineup, there are two outcomes in which there is uncertainty about

⁴ Participants who identified two individuals from the sequential lineup procedure provided a single justification statement after viewing all the remaining lineup members.

Table 1
Lineup Decisions as a Function of Lineup Presentation Format, Distance, and Culprit Presence

Lineup presentation format													
		Simultaneous					Sequential						
Culprit presence	Distance (ft)	10	60	110	160	210	All	10	60	110	160	210	All
Culprit present	Suspect ID	88.46 (253)	76.00 (209)	63.33 (190)	54.81 (148)	36.68 (95)	64.39 (895)	82.09 (362)	68.96 (291)	53.01 (220)	44.10 (187)	29.67 (124)	55.85 (1,184)
	Filler ID	6.29 (18)	9.09 (25)	15.00 (45)	18.89 (51)	27.80 (72)	15.18 (211)	7.26 (32)	12.56 (53)	19.52 (81)	23.82 (101)	31.10 (130)	18.73 (397)
	Reject	5.24 (15)	14.91 (41)	21.67 (65)	26.30 (71)	35.52 (92)	20.43 (284)	10.66 (47)	18.48 (78)	27.47 (114)	32.08 (136)	39.23 (164)	25.42 (539)
Culprit absent	Suspect ID	5.08	5.75	7.23	7.54	8.21	6.75	5.83	6.14	7.42	7.45	8.50	7.02
	Filler ID	25.39 (85)	28.77 (97)	36.14 (121)	37.71 (124)	41.05 (133)	33.74 (560)	29.15 (99)	30.70 (105)	37.11 (118)	37.25 (114)	42.50 (127)	35.09 (563)
	Reject	69.53 (194)	65.48 (184)	56.63 (158)	54.74 (150)	50.74 (137)	59.51 (823)	65.02 (184)	63.16 (180)	55.47 (147)	55.29 (141)	49.00 (122)	57.89 (774)
	<i>d'</i>	2.84	2.28	1.80	1.56	1.05	1.86	2.49	2.04	1.52	1.29	0.84	1.62

Note. Values in parentheses represent the number of participants who provided that response and the values outside of parentheses are the corresponding proportions. We estimated the innocent-suspect identification (ID) rate by dividing the total culprit-absent false-alarm rate by six (the number of lineup members) and estimated the culprit-absent filler ID rate by multiplying the total culprit-absent false-alarm rate by 5/6. Discriminability at each distance is indexed by d' ; see Mickes et al., 2014; A. M. Smith et al., 2021 for discussions on interpreting this metric on lineup tasks).

whether the witness made a correct or an incorrect decision. When a witness identifies a suspect, it is unclear whether that was a correct identification of a guilty suspect or an incorrect identification of an innocent suspect. Hence, the first of our two classifiers was intended to distinguish between accurate identifications of the culprit and mistaken identifications of innocent persons. Because there is no theoretical difference between the mistaken identification of an innocent suspect or filler from a culprit-absent lineup (all these individuals are innocent after all), we treated all mistaken identifications from culprit-absent lineups as false alarms. Our first classifier was trained to discriminate between culprit identifications (hits) and mistaken identifications from culprit-absent lineups (false alarms). Keeping with previous research, we refer to this as the hits-vs.-false-alarms (HFA) classifier. Likewise, when a witness rejects a lineup in actual criminal cases, it is unclear whether that witness is correctly rejecting a lineup that includes an innocent suspect (correct rejection) or mistakenly rejecting a lineup that includes the culprit (miss). Hence, our second classifier was trained to distinguish between correct rejections and misses. Keeping with previous research, we refer to this as the correct-rejections-vs.-misses (CRM) classifier. We trained separate HFA and CRM classifiers for simultaneous and sequential lineups.

Of the 6,230 participants who provided suitable data, we excluded data from those 608 participants who identified a filler in culprit-present conditions (397 from the sequential condition and 211 from the simultaneous condition)⁵ leaving us with a total of 5,622 justification statements fit for analysis; one from each participant. Of these statements, 2,562 participants viewed a simultaneous lineup and 3,060 viewed a sequential lineup. The simultaneous HFA model included a total of 1,455 justifications (895 correct identifications of the culprit and 560 culprit-absent filler identifications). The sequential HFA model included a total of 1,747 justifications (1,184 correct identifications of the culprit and 563 culprit-absent filler identifications). The simultaneous CRM model included a total of 1,107 justifications (823 correct rejections and 284 misses). Finally, the sequential CRM model included a total of 1,313 justifications (774 correct rejection and 539 misses).

Creating a classifier first involves generating a document-feature-matrix where all the words participants used to justify their decision are included as independent categorical predictor variables. Each row in the matrix represents the justification provided by a given witness, and each column represents a single discrete word from the entire collection of words participants used to justify their lineup decision. This matrix of words is commonly referred to as the “corpus.” We counted the number of times that a given word (across the corpus) was used in each participant’s justification statement and applied a binary outcome of 1 (word present) or 0 (word absent) for each cell in the matrix. The result was an entire matrix comprised of 0s and 1s (Dobbins & Kantner, 2019).

After we finalized the corpus, we trained a machine to use the word predictors to distinguish between correct decisions and incorrect decisions. We repeated this for all four models so that the machine learning classifier learned to distinguish hits from false alarms for the HFA models and correct rejections from misses for

⁵ In actual police cases there should only be one suspect in a lineup and the fillers are already known to be innocent by the police investigator. Accordingly, the utility of reflector variables is not forensically meaningful for filler identifications made in culprit-present lineup conditions.

Table 2
Discriminability Across Sequential Lineup Positions

Lineup decision and d'	Sequential lineup position					
	1	2	3	4	5	6
Culprit present suspect ID	63.32 (435/687)	50.54 (93/184)	55.68 (103/185)	57.81 (111/192)	48.48 (96/198)	51.34 (346/674)
Culprit absent filler ID	11.07 (148/1,337)	4.49 (60/1,337)	6.21 (83/1,337)	7.78 (104/1,337)	5.98 (80/1,337)	6.58 (88/1,337)
d'	1.56	1.71	1.68	1.62	1.52	1.54

Note. For the culprit present lineup condition, values in parentheses represent the number of participants who identified the culprit from the listed position divided by the number of participants who viewed the culprit in that position. For the culprit absent lineup condition, values inside parentheses represent the number of participants who identified filler from the listed position divided by the total number of witnesses who viewed a culprit-absent lineup. Values outside of parentheses are the corresponding proportions. Discriminability at each position is indexed by d' -prime (d'). ID = identification.

the CRM models. To do so, we employed a series of logistic regression analyses with a LASSO constraint (Least Absolute Shrinkage and Selection Operator; R. Tibshirani, 1996). We opted to use a LASSO constraint, because it is apt for dealing with models with numerous predictor variables where there is increased potential for model overfitting. The LASSO constraint addresses overfitting in two ways. First, the constraint appropriately shrinks large parameter estimates that might be artificially inflated due to modeling idiosyncrasies of the data set. Second, the LASSO constraint shrinks those predictor variables with near-zero parameter estimates to zero, effectively removing these variables from the model. The result of this predictor-selection property is a more parsimonious model that includes fewer predictor variables and an improved potential for the model to generalize to out-of-sample data sets (Hastie et al., 2009; James et al., 2013; R. J. Tibshirani & Wasserman, 2017).

The training procedure also determines the severity of the shrinkage operator applied to the data set. The degree to which the LASSO constrains a model's parameter estimates, λ , is determined through a 10-fold cross-validation process where a tuning parameter evaluates a range of potential λ s to arrive at one that strikes the best balance between minimizing overfitting and maintaining the model's predictive accuracy. Across 10 folds, the classifier is trained on 9/10ths of the data, which it uses to predict the remaining 1/10th of the data held out for a given fold. Through this iterative training process, the LASSO procedure not only penalizes those coefficients that are derived from noise but also removes predictors (by shrinking their coefficients to 0) that are not notably contributing to the model's predictive capacity (Hastie et al., 2009).

Figure 1 shows the HFA and CRM classifiers' training performance applied to the simultaneous models (top panel) and sequential models (bottom panel). The y-axis reflects mean deviance (classification error), which is the average tendency of the trained classifier to incorrectly classify the 10th of the data held out on each given fold at each level of λ (depicted on the x-axis), across all 10 folds of cross-validation. The deviance reflects the difference in fit between the tested model and the saturated model. A saturated model is one that perfectly fits the data; each observation is provided a prediction value and as a result the model fits the data exactly, with no residuals. Hence, greater deviance between the trained classifier's performance and the saturated model indicates a poorer model fit (Blum et al., 1999). The size of the points reflects the number of words that are retained at each level of λ . What Figure 1 shows is that imposing a constraint decreases deviance for all four models. This is evidenced by the steep decrease in slope on the left

side of the two panels. Figure 1 also shows the minimum mean cross-validation error, as indexed by the plotted triangles. In theory, this is where the model strikes the best balance between reducing overfitting and improving predictive capacity. The small increase in slope following the minimum cross-validation error locations indicate that imposing a more stringent constraint increases classification error. This is the point in the models where increasing the constraint prevents useful information from entering the model to assist in prediction. For simultaneous lineups the HFA and CRM model reached a minimum cross-validation error value by retaining 42 and 17 words at log lambdas of approximately -6 and -5 , respectively. For sequential lineups, the HFA and CRM model reached a minimum cross-validation error value by retaining 62 and 24 words, at log lambdas of approximately -6 and -5 , respectively. Had a classifier not found any words to be predictive, the triangle would have been plotted at the far right of the model's respective dotted line, indicating that 0 words were retained by the classifier.

Importantly, the classifier successfully reached a minimum mean cross-validation error for both the HFA and CRM models by retaining a subset of words as predictive. This demonstrates that across both simultaneous and sequential lineups, language can not only sort accurate from inaccurate identification decisions, but it can also sort accurate from inaccurate rejection decisions. This latter finding runs counter to the extant literature on natural language which has repeatedly demonstrated that although language is useful for postdicting the accuracy of identification decisions, it is not useful for postdicting the accuracy of rejection decisions (e.g., Dobbins & Kantner, 2019; Seale-Carlisle et al., 2022).

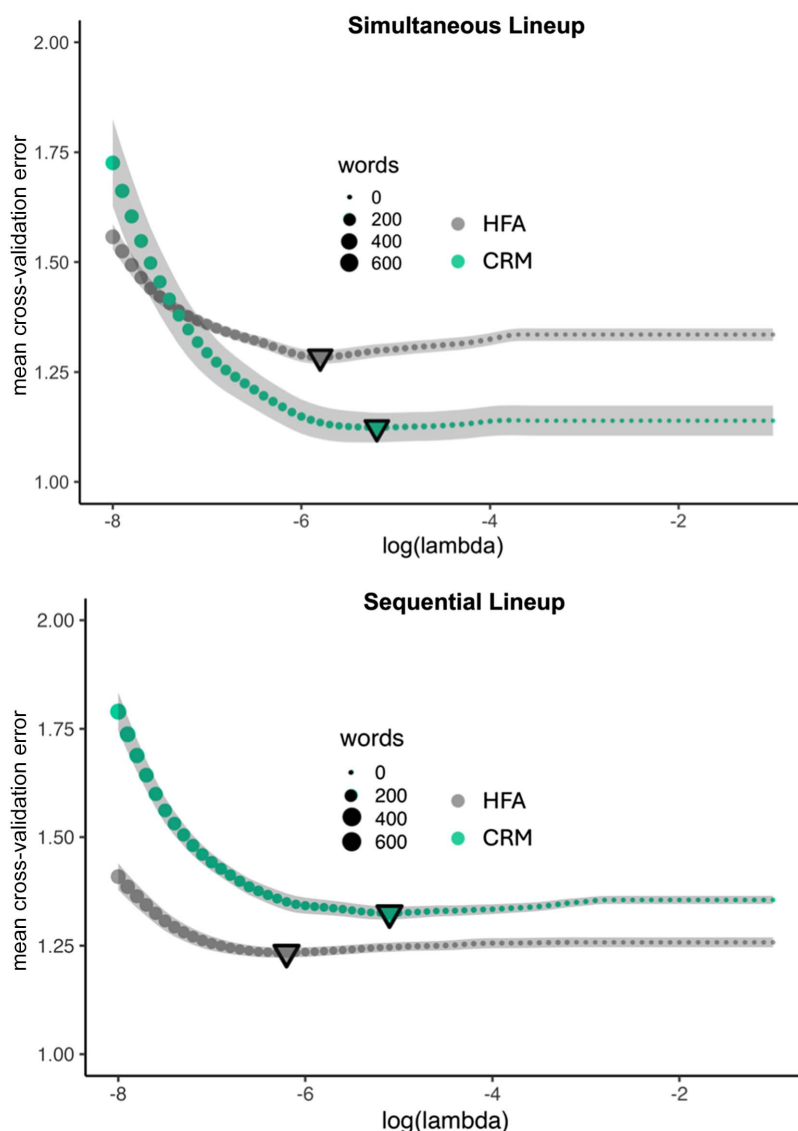
What Language Do Witnesses Use to Justify Their Lineup Decisions?

Words Diagnostic of Identification Accuracy

The retained words for identification decisions from simultaneous and sequential lineups and their regression coefficients are presented in Figures 2 and 3, respectively. For both figures the words on the left-hand side imply correct culprit identification and the words on the right-hand side imply the mistaken identification of an innocent person from a culprit-absent lineup. For both simultaneous and sequential lineups, the HFA classifiers appear to distinguish between the presence and absence of a recollective experience. Whereas hits were associated with a conscious memory experience (e.g., "recognized," "remember," "remembered") and with self-reference (e.g., "I," "my," "me"), false alarms were

Figure 1

Mean Cross-Validation Error During Training for the HFA and CRM Classifiers in the Simultaneous Lineup Condition (Top Panel) and Sequential Lineup Condition (Bottom Panel)



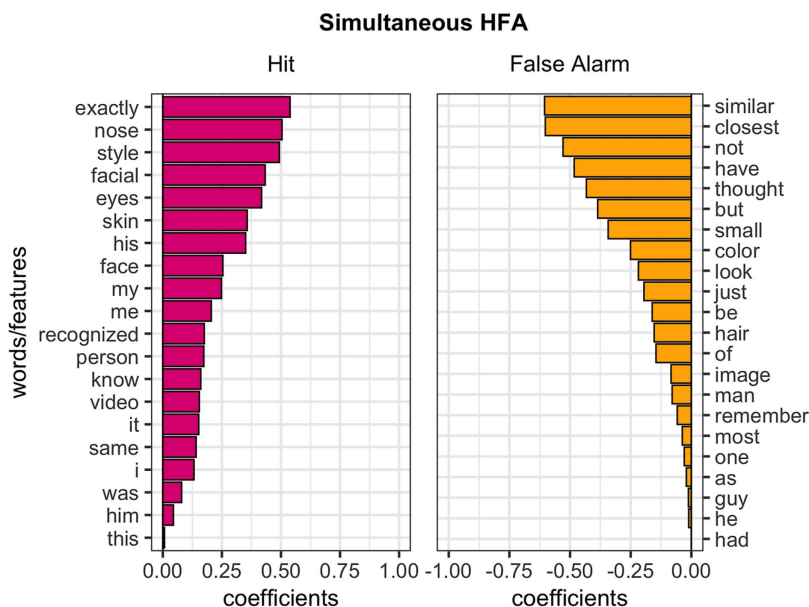
Note. The mean cross-validation error during training for the hits-vs.-false-alarms (HFA) classifier (gray) and the correct-rejections-vs.-misses (CRM) classifier (turquoise). The x -axis shows the value of the constraint on the logistic regression coefficients (λ), and the y -axis shows the mean logistic deviance (i.e., classification error) across the 10 folds of cross-validation. The size of the points indicates the number of words left in the model at different levels of constraint. The triangles indicate the minimum cross-validation mean error for each decision type. The shading reflects bootstrapped 95% confidence intervals. See the online article for the color version of this figure.

associated with equivocation (e.g., “thought,” “but,” “not,” “could”). This is remarkably consistent with the findings of Dobbins and Kantner (2019).

Consistent with Seale-Carlisle et al. (2022), witnesses mentioned observable features when justifying their identification decisions. Whereas reference to more granular features like “eyes” or “nose”

were associated with hits, reference to more coarse-grained features like “hair” or “blonde” were associated with false alarms. Although we did not predict this pattern a priori, it is consistent with the idea that witnesses who make correct identifications have stronger and more vivid memories for the culprit than do witnesses who make false identifications.

Figure 2
Retained Words From the Simultaneous HFA Classifier



Note. Figure displaying the classifier logistic lasso solution for eyewitnesses who made identification decisions from simultaneous lineups. The left side of the figure illustrates those words that are predictive of a correct identification of the culprit, while the right side of the figure illustrates those words that are predictive of a mistaken identification from a culprit-absent lineup. The words are ordered from most predictive (top) to least predictive (bottom) as indicated by the coefficient value listed on the x-axis. HFA = hits-vs.-false-alarms. See the online article for the color version of this figure.

Stepping back a bit further, the HFA classifier more generally appears to recover a distinction between absolute and relative decision processes. Not only were hits associated with words that suggest a conscious memory experience, but also with language that suggests absolute comparison of the identified lineup member to memory (e.g., “nose,” “facial,” “eyes,” “skin,” “face”) and with decisive language (e.g., “exactly,” “know,” “same,” “matched”). Comparatively, false alarms were associated with language that suggests a more relative process (e.g., “similar,” “closest,” “seems,” “most,” “like,” “familiar”) and with equivocation. This is consistent with past work in which witnesses were asked to self-report on their decision strategies (Dunning & Stern, 1994).

Notably, there appears to be little difference between simultaneous and sequential lineups in the postdictive utility of absolute versus relative language. For both simultaneous and sequential lineups, culprit identifications were associated with absolute language and mistaken identifications with relative language.

Putting the Language of Hits and False Alarms in Context. To better understand why certain words are associated with accuracy and others with inaccuracy, it is useful to examine language in the context of intact witness justifications. The HFA classifier assigns each witness’ justification a log-odds score that reflects the likelihood that justification is associated with a hit rather than false alarm. Below are the statements for which the classifiers rated as most likely to be accurate. Words that were retained as predictive of hits are bolded and words that were retained as predictive of false alarms are italicized.

Very high likelihood of a hit (simultaneous lineup condition)

1. **i remembered his face and his features i was able to study his eyes skin** tone and haircut carefully **i spent the last of my time looking at his mouth ears and facial hair**
2. **i am fairly certain i recognized the facial features of this person** the distance between **his eyes was** slightly greater than usual and reminded **me of** a teacher **i once had** so when **i saw the same features** in the lineup **i was pretty sure it was him**

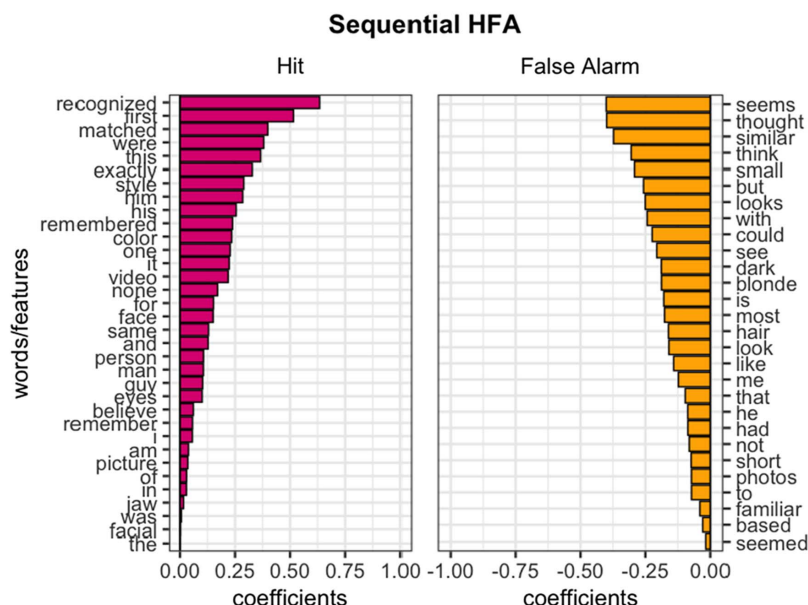
Very high likelihood of a hit (sequential lineup condition)

1. **it was the first one i recognized his face the others were definitely not him**
2. **i only identified the first photo not the secondary set none in the secondary set looked like the person in the video only the first image i was shown resembled the person in the video as to why it looked like the person i saw in the video**

These statements seem to reflect an absolute match-to-memory experience between the witness’ memory for the culprit and a lineup member. Now consider the statements that the HFA classifiers rated as most likely to be false alarms.

Very high likelihood of a false alarm (simultaneous lineup condition)

Figure 3
Retained Words From the Sequential HFA Classifier



Note. Figure displaying the classifier logistic lasso solution for eyewitnesses who made identification decisions from sequential lineups. The left side of the figure illustrates those words that are predictive of a correct identification of the culprit, while the right side of the figure illustrates those words that are predictive of a mistaken identification from a culprit-absent lineup. The words are ordered from most predictive (top) to least predictive (bottom) as indicated by the coefficient value listed on the *x*-axis. HFA = hits-vs.-false-alarms. See the online article for the color version of this figure.

1. **i** felt that the overall general *look of* the **person was similar** to the blurred picture shown to **me** the *hair color* matched *as well as* the head tilt **i** do *not* think any other photo in the line up **was** giving *similar* clues
2. the *man* had *similar hair* close cropped light *color* and a *similar* head shape

Very high likelihood of a false alarm (sequential lineup condition)

1. **the picture was** extremely *small* so **i** *could not* be sure *but i* tried to focus on **the hair it was the** only thing **i** *could kind of see but* nothing **was** really clear because **the picture was** so *small*
2. ahh **it seems** my memory is *not* as good as **i** *thought seems* **the most like the one in** my mind

These statements seem to capture witnesses' weak memory and reliance on the lineup member's hair for making an identification decision.

Words Diagnostic of Rejection Accuracy

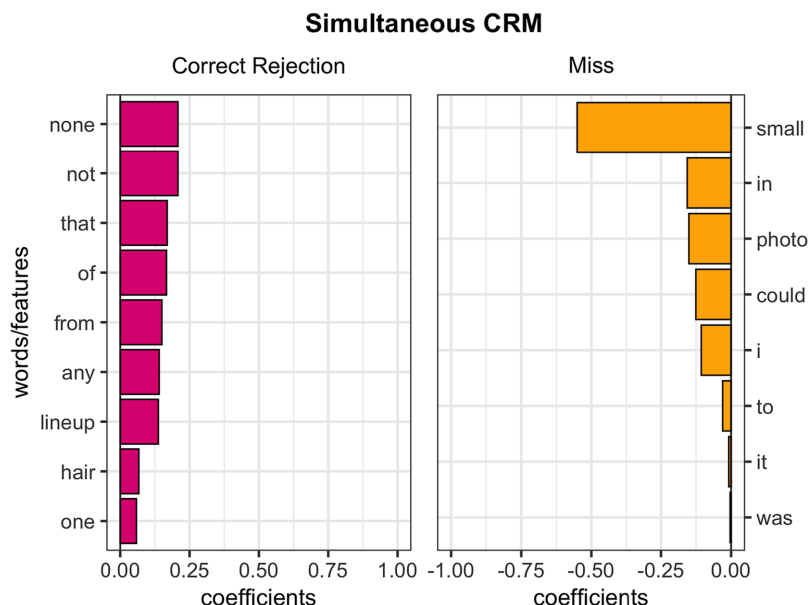
The words that the classifiers retained for rejection decisions made from simultaneous and sequential lineups and their respective regression coefficients are presented in Figures 4 and 5, respectfully. Those words presented on the left-hand side of both figures are words that imply a correct rejection and the words on the right-hand

side imply a miss. For simultaneous lineups, the CRM model appears to recover a distinction between certainty and uncertainty of the culprit's presence. Correct rejections were associated with words that suggest a decisive or absolute rejection of the lineup (e.g., "none," "not," and "any"). Misses were associated with words that suggest weak memory ("small") or uncertainty ("could"). Note that "small" is a reference to the size of the encoding video which we manipulated as a proxy for viewing distance.

For sequential lineups, it is less clear what the CRM model is distinguishing between. Though the words associated with misses, like "small" and "could," suggest weak memory and equivocation, we cannot make heads or tails of the words associated with correct rejections. To improve our ability to interpret the theme underlying correct rejection decisions from sequential lineup procedures, we built a trigram classifier model comprised of three-word-long predictors (see Figure 6). The process for building a trigram classifier is the same as that of the unigram (single-word) model. This trigram model largely captures the themes of the simultaneous CRM unigram model; correct rejections were associated with an absolute rejection of the lineup (e.g., "none of the," "not look like"), and misses were associated with weak memory for the culprit and task difficulty more generally (e.g., "hard to see," "was very small").

Putting the Language of Rejections and Misses in Context. We now evaluate the intact witness justifications provided for rejection decisions to better understand why certain words are associated with accuracy and others with inaccuracy. The CRM classifier assigns each witness' justification a log-odds score

Figure 4
Retained Words From the Simultaneous CRM Classifier



Note. Figure displaying the classifier logistic lasso solution for eyewitnesses who made rejection decisions from simultaneous lineups. The left side of the figure illustrates those words that are predictive of a correct rejection, while the right side of the figure illustrates those words that are predictive of a miss. The words are ordered from most predictive (top) to least predictive (bottom) as indicated by the coefficient value listed on the x-axis. CRM = correct-rejections-vs.-misses. See the online article for the color version of this figure.

that reflects the likelihood that justification is associated with a correct rejection rather than a miss. Below are the statements for which the classifiers rated as most likely to be accurate. Words that were retained as predictive of correct rejections are bolded and words that were retained as predictive of misses are italicized.

Very high likelihood of a correct rejection (simultaneous lineup condition)

1. the shape **of** the persons face and the features did **not** match **any of** the photos the **hair** also did **not** match though *i* did consider **that** *could* be changed but shape **of** face and features are harder *to* disguise **not that** disguise *was* the intent
2. the picture *in* the top right corner **of** the **lineup** did look somewhat similar *to* the man *in* the video but *it* did **not** really match my memory **of that** face *i* thought *it was* better *to* say **that** *i* did **not** fully recognize **any of** the faces than *to* somewhat knowingly select **one of** the faces incorrectly

Very high likelihood of a correct rejection (sequential lineup condition [trigram model])⁶

1. *i* tried to pay close attention to **the man in the video** the image was small and blurry to start then got clearer toward the end of the video **the man in the video** did not have a particularly unique face shape but the skin tone hair color were close in most of the pictures shown **none**

of the pictures however looked like **the man in the video** to me

2. *i* do not believe the person **in the video** was present in the photos they all appeared to be white males with light or lightish hair as the person was **in the video** but **none of** the facial features seemed to match the person **in the video**

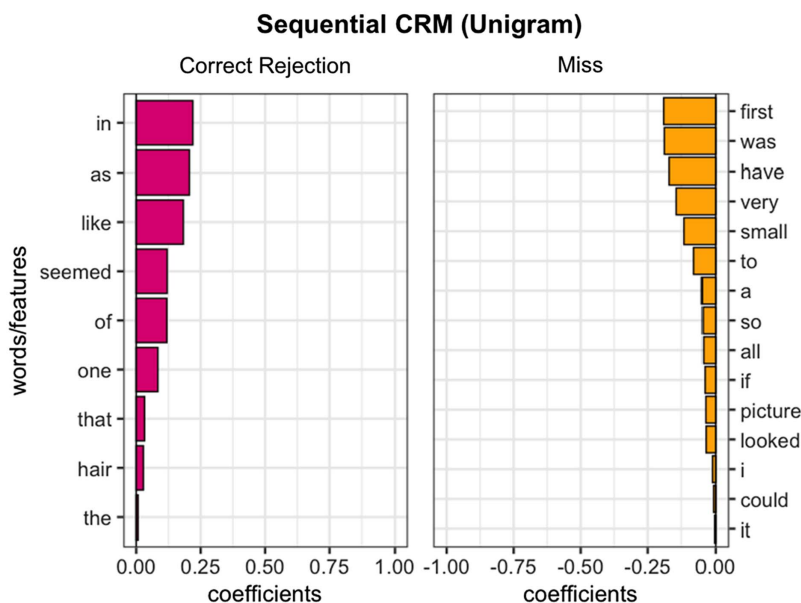
These statements appear to reflect a recall-to-reject experience. Witnesses recognize the absence of the culprit from the lineup and reject the lineup for this reason. Now consider the statements that the CRM classifiers rated as most likely to be misses.

Very high likelihood of a miss (simultaneous lineup condition)

1. although the *photo* was very *small* *i* felt the head shape was different than those *in* the *in* the line up
2. the image *i* got *in* the video was too *small to* see **any** facial features there were two photos *in* the set *i* thought *could* possibly be correct

⁶ The two statements that the sequential unigram CRM model rated as most likely to be accurate were as follows: (1) **the guy in the video looked like a walmart version of paul rudd** nobody in these set of pictures looked anything like him **as a matter of fact one looked like the illegitimate son of bruce campbell** (2) **one of the distinct features of the man in the video was his hair** no one in the set of pictures had **the same hair as the man in the video**.

Figure 5
Retained Words From the Sequential CRM Classifier (Unigram Model)



Note. Figure displaying the classifier logistic lasso solution for eyewitnesses who made rejection decisions from sequential lineups (unigram model). The left side of the figure illustrates those words that are predictive of a correct rejection, while the right side of the figure illustrates those words that are predictive of a miss. The words are ordered from most predictive (top) to least predictive (bottom) as indicated by the coefficient value listed on the x-axis. CRM = correct-rejections-vs.-misses. See the online article for the color version of this figure.

Very high likelihood of a miss (sequential lineup condition [trigram model])⁷

1. *i was not* sure about number two and should have identified him even so *i was not* one hundred percent sure
2. there were some that i thought could be the same man but *i was not* sure and if it was a police lineup i would not want to pick somebody if *i was not* sure

Witnesses' references to poor viewing conditions suggest that misses are potentially attributable to weak memory for the culprit. Interestingly, it also appears that witnesses at some point during the lineup considered identifying someone but ultimately opted to reject.

Language Classifier Accuracy

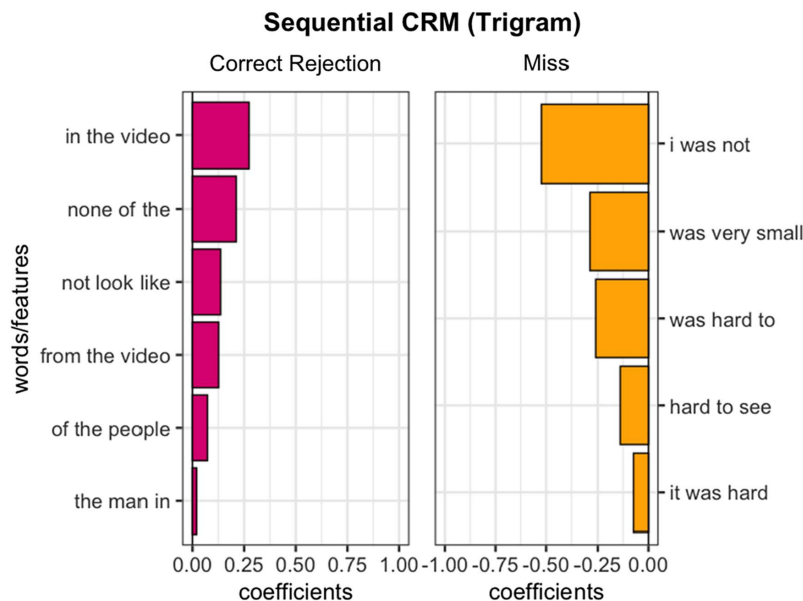
So far, we have demonstrated that machine learning classifiers recover distinct patterns of natural language for accurate and inaccurate lineup decisions. However, just how accurate are these classifiers at discriminating accuracy on the basis of language? To investigate this question, we used receiver operating characteristic analysis (ROC; Macmillan & Creelman, 2005).

The ROC classifier plots sensitivity (the true positive rate) against the false positive rate (1—specificity [the true negative rate]) at increasingly liberal classification thresholds. For HFA classifiers, these classification thresholds are based on the log odds that the

justification statement is from the hit class (indexed by positive log odds values) or the false-alarm class (indexed by negative log odds values). For the CRM classifiers, these classification thresholds are based on the log odds that the justification statement is from the correct rejection class (indexed by positive log odds values) or the miss class (indexed by negative log odds values). The most stringent classification threshold would be the highest log odds value produced by the model and should result in a high proportion of correct classifications (hits for the HFA classifier and correct rejections for the CRM classifier). This is the first point plotted on an ROC curve and this pattern continues until the most lenient classification threshold is plotted. These points are then connected and the area under the curve (AUC) represents classifier accuracy. More specifically, the AUC reflects the probability that the model will assign a randomly chosen positive case higher log odds than a randomly chosen negative case. Perfect classification performance

⁷ The two statements that the sequential unigram CRM model rated as most likely to be inaccurate were as follows: (1) **the first** participant *looked* close but almost *looked* rendered or **as if** it were a computer generated version of **the person** *the* video *i* had *was small* and *very blurry* until **the** last second **of the** video but *i* feel **like** *i* got *a* good enough look **that** **the first** picture shown *was not* him *i* am *not very* confident **that** *it was not* him however and *could have* simply been an approximately close enough resemblance *to have* been him (2) *i* think **the first** photo *i* saw might *have* been him now **that** *i have* seen **all the** photos *i* thought **that** *i* would be presented *very very* similar looking blonde **hair** males but they appeared *to be very* different *so if i could* do it again *i* would choose **the first** photo.

Figure 6
Retained Words From the Sequential CRM Classifier (Trigram Model)



Note. Figure displaying the classifier logistic lasso solution for eyewitnesses who made rejection decisions from sequential lineups (trigram model). The left side of the figure illustrates those words that are predictive of a correct rejection, while the right side of the figure illustrates those words that are predictive of a miss. The words are ordered from most predictive (top) to least predictive (bottom) as indicated by the coefficient value listed on the x-axis. CRM = correct-rejections-vs.-misses. See the online article for the color version of this figure.

would result in an AUC of 1, and at chance performance would result in an AUC of .5. Figure 7 shows the ROC curve for the HFA and CRM classifiers for both sequential and simultaneous lineups.⁸ In all cases, the classifiers performed above chance. The AUC values ranged from 0.64 for the sequential lineup CRM classifier to 0.71 for the simultaneous lineup HFA classifier. In other words, there is a 64% chance that the sequential lineup CRM classifier will assign a randomly chosen correct rejection justification a higher log odds than a randomly chosen miss justification, and that there is a 71% chance that the simultaneous lineup HFA classifier will assign a randomly chosen hit justification a higher log odds than a randomly chosen false-alarm justification.

That language is useful for sorting reliable from unreliable rejection decisions is fascinating considering that past work has only found language useful for discriminating hits from false alarms and has consistently found language unable to sort correct rejections from misses (Dobbins & Kantner, 2019; Seale-Carlisle et al., 2022). We revisit discrepancies between past and present work in the Discussion section.

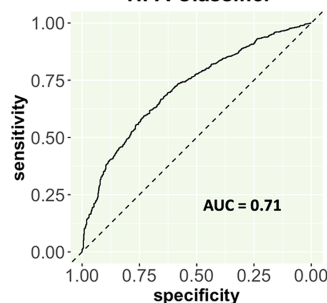
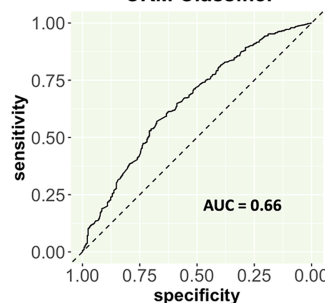
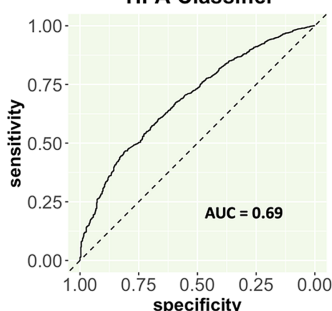
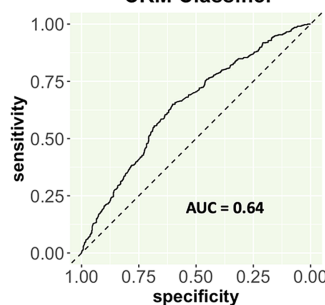
Can a Multiple-Reflector-Variable Approach to Postdicting Improve Classification Accuracy?

Next, we examined whether decision time and language increased classification accuracy over and above confidence. To evaluate the incremental validity of confidence, decision time, and language, we fit a series of (nonregularized) hierarchical logistic regression models. We started with four base models—simultaneous HFA,

simultaneous CRM, sequential HFA, and sequential CRM—and for each model, we regressed culprit-presence on witness confidence (Block 1).⁹ For each of these four models, we saved the model-predicted log odds of culprit-presence for each observation and used this variable as the confidence-only classifier in subsequent ROC

⁸ As with the broader literature on recognition memory, ROC analysis in the context of eyewitness identification typically conditions witness responding on ground truth. The probability of a response given culprit presence is plotted on the vertical axis, and the probability of a response given culprit absence is plotted on the horizontal axis. The focus is on classifying stimuli (suspect guilt status), which is sometimes referred to as a type 1 signal detection task. The type 1 signal detection task can be juxtaposed with a type 2 signal detection task, where the focus is on classifying accuracy. The type 2 ROC curve conditions witness confidence on witness accuracy. The probability of observing a confidence rating given that the witness made an accurate decision is plotted on the vertical axis and the probability of observing a confidence rating given that the witness made an inaccurate decision is plotted on the horizontal axis (e.g., Fleming & Lau, 2014; Maniscalco & Lau, 2014). The ROC curves depicted throughout this article are unique in that they can be interpreted as either type 1 or type 2 ROC curves. This is a direct result of building separate ROC curves for HFA and CRM models. For the HFA model, the vertical axis represents both culprit presence and a correct decision, and the horizontal axis represents both culprit absence and an incorrect decision. For the CRM model, the vertical axis represents both culprit absence and a correct decision, and the horizontal axis represents both culprit presence and an incorrect decision.

⁹ Note that for the HFA models, a hit corresponds to culprit presence (correct response) and a false alarm (incorrect response) corresponds to culprit absence. For the CRM models, a correct rejection corresponds to culprit absence (correct response) and a miss corresponds to culprit presence (incorrect response).

Figure 7*Training ROCs for Simultaneous and Sequential Lineups' HFA and CRM Language Classifiers***(A) Training ROC for Simultaneous Lineup HFA Classifier****(B) Training ROC for Simultaneous Lineup CRM Classifier****(C) Training ROC for Sequential Lineup HFA Classifier****(D) Training ROC for Sequential Lineup CRM Classifier**

Note. ROC = receiver operating characteristic; AUC = area under the curve; HFA = hits-vs.-false-alarms; CRM = correct-rejections-vs.-misses. See the online article for the color version of this figure.

analyses. On Block 2, we added decision time to each of the four models, so that now culprit presence was being regressed on both confidence and decision time. For each of the four models, we once again saved the model-predicted log odds of culprit-presence for each observation and used this variable as a classifier in subsequent ROC analyses. Finally, on Block 3, we regressed culprit presence on confidence, decision time, and the log-odds estimates provided by the language classifiers that were used to generate the ROC curves depicted in Figure 7. Then for each of the models (simultaneous HFA, simultaneous CRM, sequential HFA, and sequential CRM), we assessed incremental validity by determining whether the classifier based on both confidence and decision time had a larger AUC than the classifier based only on confidence, and by determining whether the classifier based on confidence, decision time, and language, had a larger AUC than the classifier based on only confidence and decision time.

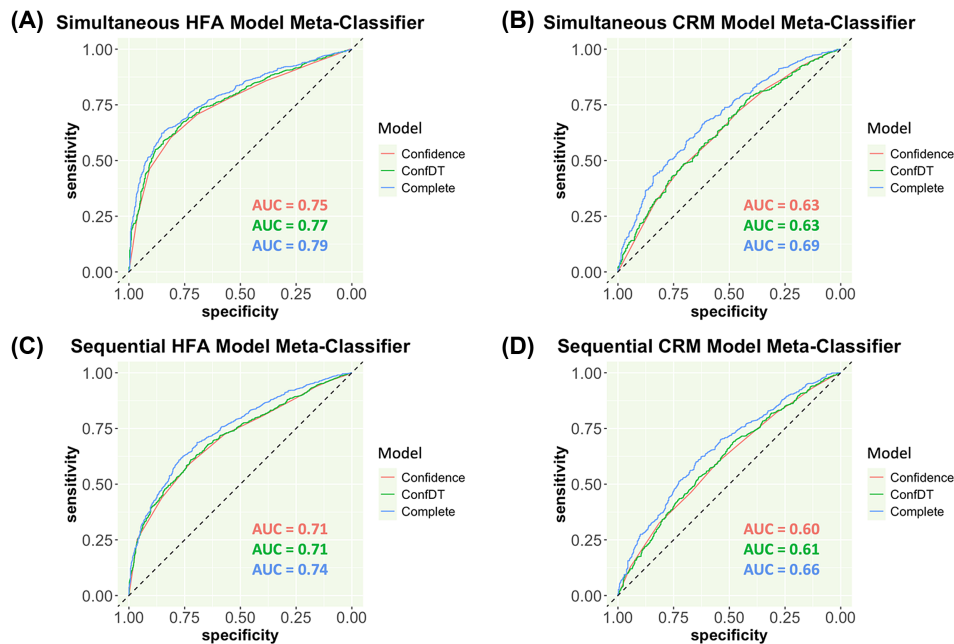
Compared to relying only on confidence, using decision time and confidence together increased the capacity to discriminate hits from false alarms for simultaneous lineups, $D = 4.00$, $p < .001$ (confidence-only AUC = 0.75, confidence-plus-decision-time AUC = 0.77), but not notably so for sequential lineups, $D = 2.08$, $p = .04$ (confidence-only AUC = 0.71, confidence-plus-decision-time AUC = 0.71). Adding language to confidence and decision time led to an incremental increase in discriminability for both simultaneous lineups, $D = 3.78$, $p < .001$ (AUC = 0.79), and sequential lineups, $D = 3.78$, $p < .001$ (AUC = 0.74). The respective ROC curves are depicted in Panels A and C of Figure 8.

Where the utility of language is most evident is in discriminating correct rejections from misses (see Panels B and D of Figure 8). For both simultaneous and sequential lineup conditions, using decision time and confidence together did not lead to an improvement in discriminability compared to just using confidence in isolation (simultaneous lineup: $D = 0.92$, $p = .36$, confidence-only AUC = 0.63, confidence-plus-decision-time AUC = 0.63); sequential lineup: $D = 0.99$, $p = .32$, confidence-only AUC = 0.60, confidence-plus-decision-time AUC = 0.61). However, adding language to confidence and decision time led to a marked improvement in discriminability for both simultaneous, $D = 3.90$, $p < .001$ (AUC = 0.69), and sequential lineups, $D = 3.58$, $p < .001$ (AUC = 0.66).

Can Language Help Flag High Confidence Errors?

At the time of the initial lineup procedure, high confidence in an identification decision implies high accuracy (Wixted & Wells, 2017). However, high confidence does not *guarantee* high accuracy (e.g., Fitzgerald et al., 2023; Giacona et al., 2021), and law enforcement personnel need mechanisms for detecting overconfidence (high-confidence errors). Likewise, there are likely instances where a witness expresses modest confidence despite having a high probability of being accurate and law enforcement personnel need mechanisms for detecting underconfidence. The language that witnesses use to justify their identification decisions might assist with detecting both overconfidence and underconfidence. To assess this possibility, we regressed culprit presence on confidence, the

Figure 8
Training Meta-Classifer Receiver Operating Characteristic Curves



Note. Incremental improvement in classifiers' capacity to discriminate hits from false alarms (Panels A and C) and correct rejections from misses (Panels B and D) for simultaneous (Panels A and B) and sequential lineups (Panels C and D). Each panel plots discriminability for three different models: a model including only confidence (in red), a model including confidence and decision time (in green), and a complete meta-classifier model including confidence, decision time, and language (in blue). HFA = hits vs. false alarms; CRM = correct rejections vs. misses; AUC = area under the curve; ConfDT = confidence and decision time model. See the online article for the color version of this figure.

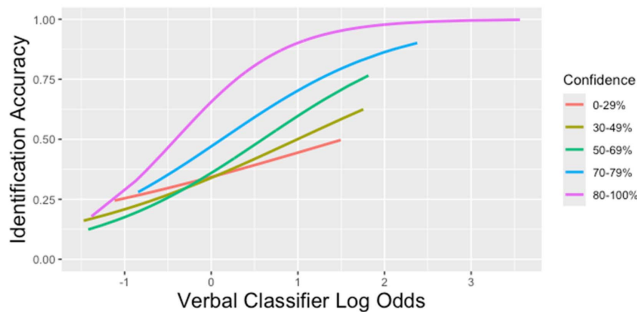
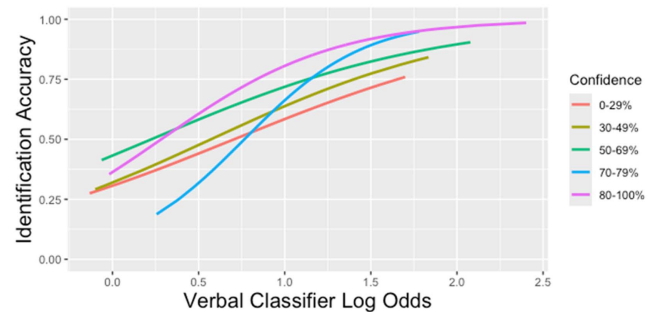
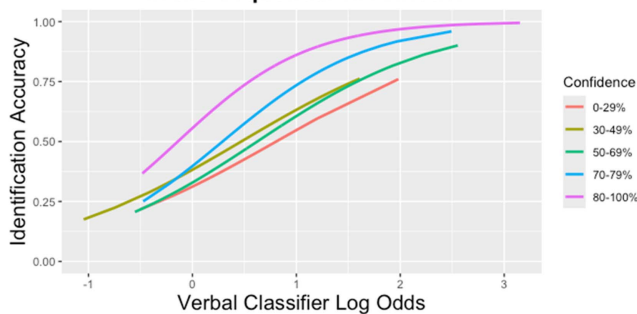
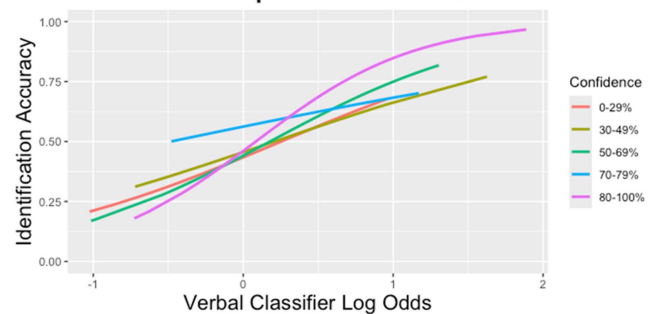
language classifier log odds, and the two-way interaction term for each of the four models: simultaneous HFA, sequential HFA, simultaneous CRM, and sequential CRM.

We found a significant interaction between the language classifier log odds and confidence for three of the four models: the simultaneous HFA classifier, $B = 0.02$, $SE = 0.005$, Wald's $z = 3.87$, $p < .001$, $e^b = 1.02$, (95% CI [1.01, 1.03]), the sequential HFA classifier, $B = 0.02$, $SE = 0.006$, Wald's $z = 2.70$, $p = .007$, $e^b = 1.02$ (95% CI [1.00, 1.03]), and the sequential CRM classifier, $B = 0.01$, $SE = 0.006$, Wald's $z = 2.07$, $p = .04$, $e^b = 1.01$ (95% CI [1.00, 1.02]). This interaction was not significant for the simultaneous CRM classifier, $B = 0.01$, $SE = 0.008$, Wald's $z = 1.91$, $p = .06$, $e^b = 1.01$, (95% CI [0.99, 1.03]; see Figure 9). Figure 9 demonstrates that language adds information above and beyond confidence that can be used to postdict accuracy. What is perhaps most noteworthy is that the language used in justifications can be used to distinguish between accurate and inaccurate high-confidence identification decisions. The easiest way to see this in Figure 9 is by considering the range of accuracy rates covered by identification decisions made with high confidence (Panels A and C). When verbal classifier scores are low, even high-confidence decisions are associated with relatively low accuracy and when verbal classifier scores are high, high confidence virtually guarantees accuracy. This suggests that the language witnesses use to justify their identification decisions is useful for determining when to trust and when not to trust a high-confidence identification decision. Likewise, Figure 9 also shows that

language might be useful for determining when a witness is underconfident. Indeed, when verbal classifier scores are high, accuracy is sometimes high even for witnesses with moderate levels of confidence.

Discussion

We examined whether using multiple reflector variables together could improve potential to distinguish between accurate and inaccurate lineup decisions on simultaneous and sequential lineup procedures. In addition to collecting confidence ratings and measuring decision time, we asked participants to justify their lineup decisions and then used a machine learning approach to determine which language best distinguished between accurate and inaccurate decisions. Confidence, decision time, and the language that witness-participants used to justify their lineup decisions all proved useful for distinguishing between accurate and inaccurate witnesses. What is more, we found evidence of incremental validity. Postdictive accuracy was maximized when all three variables were used in combination. Past research has found that confidence, decision time, and the language used to justify identification decisions is useful for distinguishing between culprit identifications and innocent-suspect identifications on simultaneous lineups. We replicated that pattern and extended it to sequential lineups. Furthermore, in contrast to previous research (e.g., Dobbins & Kantner, 2019; Seale-Carlisle et al., 2022), we found that language was apt at discriminating correct

Figure 9*Interaction Between the Language Classifier Log Odds Values and Confidence for HFA and CRM Models***(A) Classifier Log Odds and Confidence Interaction in the Simultaneous HFA Model****(B) Classifier Log Odds and Confidence Interaction in the Simultaneous CRM Model****(C) Classifier Log Odds and Confidence Interaction in the Sequential HFA Model****(D) Classifier Log Odds and Confidence Interaction in the Sequential CRM Model**

Note. Model-predicted values. The figures demonstrate how the language witnesses used to justify their lineup decisions can be used alongside confidence to better postdict lineup accuracy. HFA = hits vs. false alarms; CRM = correct rejections vs. misses. See the online article for the color version of this figure.

rejections from misses across both simultaneous and sequential lineup procedures.

The Incremental Value of Confidence, Decision Time, and Language for Sorting Reliable From Unreliable Lineup Decisions

Although many research programs have examined the potential for variables other than eyewitness confidence to postdict accuracy, few have examined the incremental validity that comes from using multiple reflector variables in concert (e.g., N. Brewer & Wells, 2006; Dodson & Dobolyi, 2016; Grabman et al., 2019; Seale-Carlisle et al., 2022; Weber et al., 2004). Of those that have looked at combining reflector variables, the primary focus has been on confidence and decision time and indeed, past work suggests that using those two reflector variables together increases the potential to postdict accuracy compared to using each on its own. More recent work has found that the language used to justify confidence ratings further increases the ability to postdict accuracy over and above numeric confidence and decision time alone (Seale-Carlisle et al., 2022) and even more, that the language that witnesses use to justify their identification decisions might be a better postdictor than decision time (Dobbins & Kantner, 2019). In line with these past findings, we found that when used in combination, confidence, decision time, and language worked better together than they did in

isolation. However, in past research, these postdictors were only able to distinguish hits from false alarms but not correct rejections from misses. Yet, here, we found that confidence, decision time, and decision justifications all distinguished between correct rejections and misses, and that they worked better in combination than in isolation.

Why did we find that reflector variables can distinguish between correct rejections and misses even though past studies have failed to find this effect? At the outset of this experiment, we reasoned that the failure to find that reflector variables could distinguish between correct rejections and misses might have been attributable to a restriction of range issue (D. S. Lindsay et al., 1998, 2000; Molinaro et al., 2021; Perfect, 2002). This is one reason why we decided to vary the quality of encoding conditions across participants. In our [Supplemental Materials](#) document, we present a set of analyses that assessed whether the potential to distinguish correct rejections from misses depended on heterogeneity of encoding conditions. We specifically examined whether the association between confidence and accuracy was weaker within distance conditions compared to when we collapsed across distance conditions. It was not. The only exception to this rule was in the extreme condition where witnesses were given a very poor view of the culprit and demonstrated very poor discriminability. In this situation, witnesses had very low accuracy and confidence was not significantly related to accuracy. Aside from this extreme condition, confidence consistently discriminated between correct rejections and misses.

The question remains then, why did confidence, decision time, and decision justifications distinguish correct rejections from misses despite failing to do so in past studies (e.g., Dobbins & Kantner, 2019; Seale-Carlisle et al., 2022)? We can only speculate, but one reason might be that these past studies used repeated-measures designs. In each of these studies, participants were presented with a series of targets to study and then after a short delay completed a series of test trials. When a target stood out as providing a relatively strong match to memory, participants presumably interpreted that as evidence that the target had been previously studied. But on trials in which a participant had a relatively weak recognition experience, it might have been less clear as to whether that weak recognition experience was attributable to target absence or to weak memory. Indeed, on target-absent trials there likely would have been some ambiguity surrounding which memory trace was being probed by the test items. After all, if a participant studied a relatively long list of items and had a relatively weak recognition experience at the time of test, then how would that person know which memory trace was being probed? Conversely, in the present work, each witness-participant saw only one target individual and then completed either a culprit-present or culprit-absent lineup. Witnesses had only one trace from the encoding context, and therefore, even if they had a weak recognition experience, they clearly knew which trace was being probed at the time of the lineup. This should have led to a situation where witnesses with relatively strong memory traces were relatively accurate and displayed behaviors that implied accuracy and witnesses with relatively weak memory traces were relatively inaccurate and displayed behaviors that implied inaccuracy. If our speculation is correct, we are encouraged that the current findings might better characterize eyewitness identification in actual cases, which tend to not have the repeated-measures characteristic.

What Language Distinguishes Accurate From Inaccurate Eyewitnesses?

Consistent with previous research, we found that culprit identifications were associated with references to recollection and mistaken identifications were associated with references to equivocation (Dobbins & Kantner, 2019; Seale-Carlisle et al., 2022). We also found that whereas culprit identifications were associated with use of absolute language, mistaken identifications were associated with use of relative language. Indeed, witnesses who identified the culprit used words like “exactly,” “matched,” or “recognized” and witnesses who mistakenly identified innocent persons used words like “similar,” “closest,” or “most.” Perhaps this association between absolute language and decision accuracy should be expected given the association between decision accuracy and recollection. Indeed, a recollective experience is a conscious and vivid recognition experience, and it is only natural that such an experience would be associated with use of more absolute language. Conversely, in the absence of recollective experience, witnesses might be forced to fall back on secondary processes such as a relative-judgment strategy (Charman & Wells, 2007; A. M. Smith, Ying, Goldstein et al., 2024). Several previous experiments have found that whereas accurate witnesses report using absolute-judgment strategies, inaccurate witnesses report using relative-judgment strategies (Dunning & Perretta, 2002; Dunning & Stern, 1994; Kneller et al., 2001; Wells, 1984, 1993).

It is important that we emphasize the distinction between affecting accuracy and reflecting accuracy. The present findings do not suggest that using an absolute-judgment strategy leads to accurate identification decisions and using a relative-judgment strategy leads to inaccurate identification decisions. What they do imply is that absolute-judgment strategies REFLECT accuracy and relative-judgment strategies REFLECT inaccuracy. Why might this be? As memory strength increases, lineup tasks become easier, and witnesses become more likely to make accurate identification decisions (Charman & Wells, 2007). If a task is easy, then there would be little reason for a witness to engage in an effortful and deliberative judgment process, and we might expect that witness to justify her identification decisions with absolute language. Conversely, a witness with a weaker memory would find the task more difficult and might be forced to fall back on a relative-judgment strategy to determine whether the culprit is present in the lineup and to determine who that is. Hence, both eyewitness identification accuracy and the language that witnesses use to justify their identification decisions are likely caused by variations in memory strength. In fact, confidence and decision time should also be caused, at least in part, by memory strength. As memory strength increases, witnesses become increasingly likely to make accurate identification decisions, they make their decisions more quickly, and they express greater confidence in those decisions.

Correct rejections were also associated with use of absolute language, on both simultaneous and sequential lineups. The language that witnesses used to justify correct rejections was consistent with use of a recall-to-reject strategy (e.g., Brown et al., 1977; Gallo, 2004; Ghetti, 2003; Rotello et al., 2000; Strack & Bless, 1994). A recall-to-reject strategy (i.e., *recollection rejection*) refers to a process where an individual can vividly recall aspects of the encoding situation that are discrepant with the test probe and uses this discrepancy as a basis for rejection (e.g., Brainerd et al., 2003; Hintzman & Curran, 1994; Lampinen & Odegard, 2006; Odegard & Lampinen, 2005). In the context of an eyewitness lineup, a witness might vividly recall the culprit’s face, detect a discrepancy between the culprit’s face and the faces presented on the lineup, and use this as a basis for rejection. In the present work, correct rejections of simultaneous lineups were associated with absolute language like “not,” “none,” “any,” and correct rejections of sequential lineups were associated with language like “none of the” and “not look like.” Although we cannot definitively tell from these justifications whether witnesses engaged in recollection rejection, the language used is consistent with that process.

Why Do Reflector Variables Do a Better Job at Postdicting Suspect Identifications Than Lineup Rejections?

Since collecting the data for this experiment, we have determined that a major contributing factor for why reflector variables better postdict the accuracy of suspect identifications than the accuracy of rejection decisions has to do with how lineups are designed. When a witness identifies a police suspect, expressions of confidence (and other witness behaviors) are directly yoked to that suspect and directly assess the degree of match between that person and the witness’ memory for the culprit. This serves as an excellent proxy for the likely guilt of the suspect. But when a witness rejects a lineup,

the filler will typically provide a stronger match to the witness' memory for the culprit than will the suspect. In that case, confidence and other reflector variables do not measure the degree of match between the suspect and the witness' memory for the culprit (A. M. Smith & Ayala, 2021; Starns et al., 2023; Yilmaz et al., 2022). Instead, reflector variables index the degree of match between the filler and the witness' memory for the culprit, which does not inform on the likely guilt or innocence of the suspect. Instead, of using a standard lineup procedure where a witness makes a lineup decision and provides only a single confidence rating, lineups should be modified to ensure that no matter which lineup member provides the strongest match to the witness' memory for the culprit, there is always information that directly scales the degree of match between the suspect and the witness' memory for the culprit. When witnesses who reject lineups are asked to assign a confidence rating to each lineup member, confidence is strongly related to the accuracy of rejection decisions (Ayala et al., 2022; A. M. Smith et al., 2023; A. M. Smith, Ying, Ayala et al., 2024). In short, a major reason for the asymmetric utility of reflector variables is that the traditional lineup often does not measure the degree of match between the suspect and the witness' memory for the culprit.

Comparing Simultaneous Lineups and Sequential Lineups

The language that witnesses used to justify their lineup decisions was similar for both simultaneous and sequential lineups. Across both presentation types, accurate witnesses used more absolute language and inaccurate witnesses used more relative language. The present data show no evidence that witnesses presented with sequential lineups were more likely to engage in absolute-judgment strategies or less likely to engage in relative-judgment strategies than were witnesses presented with simultaneous lineups. To more formally test this proposition, our [Supplemental Materials](#) present a series of analyses in which we used machine learning classifiers to distinguish between witnesses who completed simultaneous lineups and witnesses who completed sequential lineups. Although the classifiers were able to distinguish between simultaneous and sequential lineups, the classifiers do not appear to be recovering a difference between absolute- and relative-judgment strategies. The sequential lineup procedure was designed with the goal of making it more difficult for witnesses to compare faces and to engage in relative-judgment strategies (R. C. L. Lindsay & Wells, 1985; Steblay et al., 2011). Undoubtedly, a sequential lineup would make the relative comparison process more difficult. After all, the sequential procedure does not permit witnesses to examine more than one face at a time. But making relative-judgment strategies more difficult and decreasing the use of relative-judgment strategies are two different things. The present results suggest that witnesses who mistakenly identified innocent persons from culprit-absent sequential lineups used the same types of relative language as did witnesses who mistakenly identified innocent persons from culprit-absent simultaneous lineups.

It is possible that variation in memory strength exerts such a strong force on judgment strategy that it outshines most any attempt to influence judgment strategy through variation in procedure. There are numerous examples of researchers attempting to encourage use of absolute strategies or discourage use of relative strategies with little effect. Several researchers have attempted to decrease use of

relative-judgment strategies by placing limits on the amount of time witnesses have to make lineup decisions. None of these attempts affected accuracy (N. Brewer et al., 2000; Charman, 2004; Perretta & Dunning, 2001). Instructing witnesses to use absolute- or relative-judgment strategies also had no impact on eyewitness decision-making (Moreland & Clark, 2020). Perhaps these manipulations were unsuccessful because they pale in comparison to the impact of memory strength. Even the most well-intentioned of witness-participants would be unlikely to engage in a relative-judgment strategy if one of the lineup members just popped out. Likewise, if a witness has trouble even forming a preference for a lineup member, it might be difficult to avoid engaging in the sorts of comparative processes that typify relative-judgment strategies.

Finally, we must question whether attempts to increase identification accuracy by manipulating judgment strategies were misguided. Consider our earlier point that judgment strategies do not AFFECT accuracy but instead REFLECT accuracy (A. M. Smith & Wells, 2023; Wells, 2020). From this perspective, the reason judgment strategy is associated with identification accuracy is because both are effects of memory strength (Charman & Wells, 2007). As memory strength increases, witnesses are both more likely to adopt absolute-judgment strategies and to make accurate decisions. As memory strength decreases, witnesses are both more likely to adopt relative-judgment strategies and to make inaccurate decisions. But if both judgment strategy and accuracy are caused by memory strength and if judgment strategy does not affect accuracy, then why should manipulations of judgment strategy affect accuracy? It would seem to us that they should not and that might explain why the sequential lineup did not lead to superior accuracy in the present work.

Although the purpose of this article was not to compare the discriminability of simultaneous and sequential lineups, it is apparent that the simultaneous lineup had better discriminability at all five viewing distances that were used in this experiment. In fact, at every single viewing distance we considered, the simultaneous lineup led to more hits and either no change or a slight decrease in false alarms. This finding adds to a growing body of evidence that simultaneous lineups better discriminate guilty suspects from innocent suspects than do sequential lineups (e.g., Dobolyi & Dodson, 2013; Finley et al., 2020; Gronlund et al., 2012; Mickes et al., 2012; Wixted et al., 2016).

Why do sequential lineups lead to worse discriminability than simultaneous lineups? One possibility is that by presenting faces to witnesses one at a time, the sequential lineup disrupts the potential for witnesses to compare faces and to determine which face provides the strongest match to their memory for the culprit (Wixted & Mickes, 2014). For example, imagine that a witness encounters the third member on a sequential lineup and that person provides a reasonably strong match to memory. Not knowing if a better option will come along, the witness identifies this individual. When the witness reaches position five, she encounters another individual who provides a reasonably strong match to memory and who unbeknownst to the witness is the actual culprit. Having already identified the person in position three, she decides to reject the person in position five. Had the lineup members been presented to the witness simultaneously, the witness might have been able to determine that the person in position five (the culprit) provided a stronger match to memory than did the person in position three (a filler) and perhaps would have correctly identified the culprit. There are alternative mechanisms that could explain the simultaneous lineup advantage, but whatever

the mechanism, it has become increasingly clear that simultaneous presentation better discriminates guilty suspects from innocent suspects than does sequential presentation (e.g., Dobolyi & Dodson, 2013; Gronlund et al., 2012; Mickes et al., 2012; Wixted et al., 2016).

How Might Multiple Reflector Variables Be Used to Evaluate Accuracy in Actual Criminal Investigations?

What is unclear up until this point is how the individuals who are charged with evaluating eyewitness accuracy should make use of multiple reflector variables. One possibility is to use machine learning models to generate an estimate of witness accuracy based on confidence, decision time, and the language that the witness used to justify her lineup decision. This actuarial information could then be provided to criminal justice actors (such as police investigators, attorneys, judges, and jurors), so that they can incorporate this information into their decision-making process (e.g., Kelso et al., 2024). Alternatively, these evaluators might opt for a heuristic approach to guide assessments of decision accuracy. For example, absolute language and language suggesting a recollective process were associated with accurate suspect identifications. Confidence and decision time were also associated with suspect identifications. Evaluators might take a unit-weighting approach to evaluating eyewitness accuracy. For example, the evaluator could rate witness language as absolute (1) or relative (0), recollective (1) or not (0), confident (1) or unconfident (0), and decision speed as fast (1) or slow (0). The scores could then be tallied to determine the likelihood that the witness made an accurate lineup decision. These so-called “improper” linear models have proven useful in numerous applied domains (Apgar, 1953; Dawes, 1979; Dawes & Corrigan, 1974; Gigerenzer & Gaissmaier, 2011). Our intention here is not to propose the model for evaluating eyewitness accuracy, but to provide an example of what a valid improper linear model might look like. One thing that we will state with confidence is that to maximize the potential of criminal justice actors to distinguish accurate from inaccurate witnesses, some sort of model will be necessary. Seventy years of research demonstrates that actuarial models consistently perform as well or better than intuitive judgment (Dawes, 1979; Grove et al., 2000; Katsikopoulos et al., 2020; Meehl, 1954). When it comes to evaluating eyewitness accuracy, the stakes are too high to simply let evaluator intuition determine which variables are relevant for assessing accuracy.

Finally, to make use of reflector variables in actual police cases, lineup procedures should be video recorded in their entirety (A. M. Smith & Wells, 2023; Wells et al., 2020). This is the only way to ensure a complete and veridical record of everything that transpires during the identification procedure. From these videos, one would be able to extract decision time, confidence, and everything the witness says and does while completing the lineup procedure. These eyewitness behaviors, or reflector variables, can then be leveraged to assess witness accuracy and suspect guilt.

Conclusion

Through machine learning analyses, we evaluated the potential for eyewitness postidentification confidence, decision time, and the language witnesses used to justify their lineup decisions to discriminate accurate from inaccurate identification and rejection

decisions from both simultaneous and sequential lineup procedures. We found that justifications that largely reflected an absolute decision-making process were indicative of accuracy. In contrast, justifications that largely reflected a relative decision-making process were indicative of inaccuracy. Finally, the results of this experiment show that the criminal justice system should consider witness confidence, decision time, and the rationale witnesses provide for their lineup decision altogether when assessing lineup accuracy. Evaluating all these behaviors in combination would improve the criminal justice system’s ability to discriminate likely accurate witnesses from likely inaccurate witnesses.

References

- Ackerman, R., & Thompson, V. (2014). Meta-reasoning: What can we learn from meta memory? In A. Feeney & V. Thompson (Eds.), *Reasoning as memory* (pp. 164–182). Psychology Press.
- Apgar, V. (1953). A proposal for a new method of evaluation of the newborn infant. *Current Researches in Anesthesia & Analgesia*, 32(4), 260–267. <https://doi.org/10.1213/00000539-195301000-00041>
- Ayala, N. T., Smith, A. M., & Wells, G. L. (2021). *Using reflector variables to sort between the presence and absence of the culprit in simultaneous and sequential lineups*. Open Science Framework. https://osf.io/x2ru4/?view_only=bcc3d6de9ccc4df380632d4ff4bc21ec
- Ayala, N. T., Smith, A. M., & Ying, R. C. (2022). The rule-out procedure: Increasing the potential for police investigators to detect suspect innocence from eyewitness lineup procedures. *Journal of Applied Research in Memory and Cognition*, 11(4), 489–499. <https://doi.org/10.1037/ma0000018>
- Baptiste, A. (2015). *gridExtra: Miscellaneous functions for “grid” graphics* (R package Version 2.0.0) [Computer software]. <https://CRAN.R-project.org/package=gridExtra>
- Behrman, B. W., & Richards, R. E. (2005). Suspect/foil identification in actual crimes and in the laboratory: A reality monitoring analysis. *Law and Human Behavior*, 29(3), 279–301. <https://doi.org/10.1007/s10979-005-3617-y>
- Benoit, K., & Obeng, A. (2021). *readtext: Import and handling for plain and formatted text files* (R package Version 0.81) [Computer software]. <https://CRAN.R-project.org/package=readtext>
- Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., & Matsuo, A. (2018). quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software*, 3(30), Article 774. <https://doi.org/10.21105/joss.00774>
- Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual Review of Psychology*, 64(1), 417–444. <https://doi.org/10.1146/annurev-psych-113011-143823>
- Blum, A., Kalai, A. T., & Langford, J. (1999). Beating the hold-out: Bounds for k-fold and progressive cross-validation. *Proceedings of the twelfth annual conference on learning theory* (pp. 203–208). Association for Computing Machinery.
- Brainerd, C. J., Reyna, V. F., Wright, R., & Mojardin, A. H. (2003). Recollection rejection: False-memory editing in children and adults. *Psychological Review*, 110(4), 762–784. <https://doi.org/10.1037/0033-295X.110.4.762>
- Brewer, N., Caon, A., Todd, C., & Weber, N. (2006). Eyewitness identification accuracy and response latency. *Law and Human Behavior*, 30(1), 31–50. <https://doi.org/10.1007/s10979-006-9002-7>
- Brewer, N., Gordon, M., & Bond, N. (2000). Effect of photoarray exposure duration on eyewitness identification accuracy and processing strategy. *Psychology, Crime & Law*, 6(1–2), 21–32. <https://doi.org/10.1080/10683160008410829>
- Brewer, N., Weber, N., Wootton, D., & Lindsay, D. S. (2012). Identifying the bad guy in a lineup using confidence judgments under deadline pressure.

- Psychological Science*, 23(10), 1208–1214. <https://doi.org/10.1177/0956797612441217>
- Brewer, N., & Wells, G. L. (2006). The confidence–accuracy relationship in eyewitness identification: Effects of lineup instructions, foil similarity, and target-absent base rates. *Journal of Experimental Psychology: Applied*, 12(1), 11–30. <https://doi.org/10.1037/1076-898X.12.1.11>
- Brewer, W. F., Sampaio, C., & Barlow, M. R. (2005). Confidence and accuracy in the recall of deceptive and nondeceptive sentences. *Journal of Memory and Language*, 52(4), 618–627. <https://doi.org/10.1016/j.jml.2005.01.017>
- Brown, J., Lewis, V. J., & Monk, A. F. (1977). Memorability, word frequency and negative recognition. *The Quarterly Journal of Experimental Psychology*, 29(3), 461–473. <https://doi.org/10.1080/14640747708400622>
- Charman, S. D. (2004). *The problem with introspective eyewitnesses: Can thinking too much harm the identification process?* [Masters thesis, Iowa State University]. Iowa State University Digital Repository. <https://dr.lib.iastate.edu/entities/publication/64f3f592-55ea-4adb-8495-4f14a1bf20d4>
- Charman, S. D., & Wells, G. L. (2007). Applied lineup theory. In R. C. L. Lindsay, D. F. Ross, J. D. Read, & M. P. Toglia (Eds.), *The handbook of eyewitness psychology: Vol. 2. Memory for people* (pp. 219–254). Lawrence Erlbaum.
- Clark, S. E., & Wells, G. L. (2008). On the diagnosticity of multiple-witness identifications. *Law and Human Behavior*, 32(5), 406–422. <https://doi.org/10.1007/s10979-007-9115-7>
- Davis, S. D., & Peterson, D. J. (2022). Simulated viewing distance impairs the confidence–accuracy relationship for long, but not moderate distances: Support for a model incorporating the role of feature ambiguity. *Cognitive Research: Principles and Implications*, 7(1), Article 55. <https://doi.org/10.1186/s41235-022-00406-5>
- Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist*, 34(7), 571–582. <https://doi.org/10.1037/0003-066X.34.7.571>
- Dawes, R. M., & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin*, 81(2), 95–106. <https://doi.org/10.1037/h0037613>
- Deffenbacher, K. A. (1980). Eyewitness accuracy and confidence: Can we infer anything about their relationship? *Law and Human Behavior*, 4(4), 243–260. <https://doi.org/10.1007/BF01040617>
- DeSoto, K. A., & Roediger, H. L., III. (2014). Positive and negative correlations between confidence and accuracy for the same events in recognition of categorized lists. *Psychological Science*, 25(3), 781–788. <https://doi.org/10.1177/0956797613516149>
- Dobbins, I. G. (2022). Recognition language classifiers demonstrate far transfer of learning. *Psychonomic Bulletin & Review*, 29(4), 1414–1425. <https://doi.org/10.3758/s13423-022-02085-1>
- Dobbins, I. G., & Kantner, J. (2019). The language of accurate recognition memory. *Cognition*, 192, Article 103988. <https://doi.org/10.1016/j.cognition.2019.05.025>
- Dobolyi, D. G., & Dodson, C. S. (2013). Eyewitness confidence in simultaneous and sequential lineups: A criterion shift account for sequential mistaken identification overconfidence. *Journal of Experimental Psychology: Applied*, 19(4), 345–357. <https://doi.org/10.1037/a0034596>
- Dodson, C. S., & Dobolyi, D. G. (2016). Confidence and eyewitness identifications: The cross-race effect, decision time and accuracy. *Applied Cognitive Psychology*, 30(1), 113–125. <https://doi.org/10.1002/acp.3178>
- Dunlosky, J., & Metcalfe, J. (2009). *Metacognition*. Sage Publications.
- Dunlosky, J., Serra, M. J., Matvey, G., & Rawson, K. A. (2005). Second-order judgments about judgments of learning. *The Journal of General Psychology*, 132(4), 335–346. <https://doi.org/10.3200/GENP.132.4.335-346>
- Dunning, D., & Perretta, S. (2002). Automaticity and eyewitness accuracy: A 10- to 12-second rule for distinguishing accurate from inaccurate positive identifications. *Journal of Applied Psychology*, 87(5), 951–962. <https://doi.org/10.1037/0021-9010.87.5.951>
- Dunning, D., & Stern, L. B. (1994). Distinguishing accurate from inaccurate eyewitness identifications via inquiries about decision processes. *Journal of Personality and Social Psychology*, 67(5), 818–835. <https://doi.org/10.1037/0022-3514.67.5.818>
- Feinerer, I., & Hornik, K. (2023). *tm: Text mining package* (R package Version 0.7-11) [Computer software]. <https://CRAN.R-project.org/package=tm>
- Feinerer, I., Hornik, K., & Meyer, D. (2008). Text mining infrastructure in R. *Journal of Statistical Software*, 25(5), 1–54. <https://doi.org/10.18637/jss.v025.i05>
- Finley, J. R., Wixted, J. T., & Roediger, H. L., III. (2020). Identifying the guilty word: Simultaneous versus sequential lineups for DRM word lists. *Memory & Cognition*, 48(6), 903–919. <https://doi.org/10.3758/s13421-020-01032-6>
- Fitzgerald, R. J., Tredoux, C. G., & Juncu, S. (2023). Estimation of eyewitness error rates in fair and biased lineups. *Law and Human Behavior*, 47(4), 463–483. <https://doi.org/10.1037/lhb0000538>
- Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*, 8, Article 443. <https://doi.org/10.3389/fnhum.2014.00443>
- Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1), 1–22. <https://doi.org/10.18637/jss.v033.i01>
- Gallo, D. A. (2004). Using recall to reduce false recognition: Diagnostic and disqualifying monitoring. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30(1), 120–128. <https://doi.org/10.1037/0278-7393.30.1.120>
- Ghetti, S. (2003). Memory for nonoccurrences: The role of metacognition. *Journal of Memory and Language*, 48(4), 722–739. [https://doi.org/10.1016/S0749-596X\(03\)00005-6](https://doi.org/10.1016/S0749-596X(03)00005-6)
- Giacona, A. M., Lampinen, J. M., & Anastasi, J. S. (2021). Estimator variables can matter even for high-confidence lineup identifications made under pristine conditions. *Law and Human Behavior*, 45(3), 256–270. <https://doi.org/10.1037/lhb0000381>
- Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology*, 62(1), 451–482. <https://doi.org/10.1146/annurev-psych-120709-145346>
- Grabman, J. H., Dobolyi, D. G., Berelovich, N. L., & Dodson, C. S. (2019). Predicting high confidence errors in eyewitness memory: The role of face recognition ability, decision-time, and justifications. *Journal of Applied Research in Memory and Cognition*, 8(2), 233–243. <https://doi.org/10.1037/h0101835>
- Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. 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(4), 221–228. <https://doi.org/10.1016/j.jarmac.2012.09.003>
- Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment*, 12(1), 19–30. <https://doi.org/10.1037/1040-3590.12.1.19>
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*. Springer. <https://doi.org/10.1007/978-0-387-84858-7>
- Hintzman, D. L., & Curran, T. (1994). Retrieval dynamics of recognition and frequency judgments: Evidence for separate processes of familiarity and recall. *Journal of Memory and Language*, 33(1), 1–18. <https://doi.org/10.1006/jmla.1994.1001>
- Ittelson, W. H. (1951). Size as a cue to distance: Static localization. *The American Journal of Psychology*, 64(1), 54–67. <https://doi.org/10.2307/1418595>
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. Springer. <https://doi.org/10.1007/978-1-4614-7138-7>

- Kassambara, A. (2023). *ggpubr: "ggplot2" based publication ready plots* (R package Version 0.6.0) [Computer software]. <https://rpkgs.datanovia.com/ggpubr/>
- Katsikopoulos, K. V., Şimşek, Ö., Buckmann, M., & Gigerenzer, G. (2020). *Classification in the wild: The science and art of transparent decision making*. MIT Press.
- Kelso, L. E., Grabman, J. H., Dobolyi, D. G., & Dodson, C. S. (2024). Does artificial intelligence (AI) assistance mitigate biased evaluations of eyewitness identifications? *Journal of Applied Research in Memory and Cognition*. Advance online publication. <https://doi.org/10.1037/ma0000192>
- Kneller, W., Memon, A., & Stevenage, S. (2001). Simultaneous and sequential lineups: Decision processes of accurate and inaccurate eyewitnesses. *Applied Cognitive Psychology*, 15(6), 659–671. <https://doi.org/10.1002/acp.739>
- Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review*, 103(3), 490–517. <https://doi.org/10.1037/0033-295X.103.3.490>
- Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software*, 28(5), 1–26. <https://doi.org/10.18637/jss.v028.i05>
- Lampinen, J. M., Erickson, W. B., Moore, K. N., & Hittson, A. (2014). Effects of distance on face recognition: Implications for eyewitness identification. *Psychonomic Bulletin & Review*, 21(6), 1489–1494. <https://doi.org/10.3758/s13423-014-0641-2>
- Lampinen, J. M., & Odegard, T. N. (2006). Memory editing mechanisms. *Memory*, 14(6), 649–654. <https://doi.org/10.1080/09658210600648407>
- Lindsay, D. S., Nilsen, E., & Read, J. D. (2000). Witnessing-condition heterogeneity and witnesses' versus investigators' confidence in the accuracy of witnesses' identification decisions. *Law and Human Behavior*, 24(6), 685–697. <https://doi.org/10.1023/A:1005504320565>
- Lindsay, D. S., Read, J. D., & Sharma, K. (1998). Accuracy and confidence in person identification: The relationship is strong when witnessing conditions vary widely. *Psychological Science*, 9(3), 215–218. <https://doi.org/10.1111/1467-9280.00041>
- Lindsay, R. C. L., Kalmet, N., Leung, J., Bertrand, M. I., Sauer, J. D., & Sauerland, M. (2013). Confidence and accuracy of lineup selections and rejections: Postdicting rejection accuracy with confidence. *Journal of Applied Research in Memory and Cognition*, 2(3), 179–184. <https://doi.org/10.1016/j.jarmac.2013.06.002>
- 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), 796–802. <https://doi.org/10.1037/0021-9010.76.6.796>
- Lindsay, R. C. L., Semmler, C., Weber, N., Brewer, N., & Lindsay, M. R. (2008). How variations in distance affect eyewitness reports and identification accuracy. *Law and Human Behavior*, 32(6), 526–535. <https://doi.org/10.1007/s10979-008-9128-x>
- Lindsay, R. C. L., & Wells, G. L. (1985). Improving eyewitness identifications from lineups: Simultaneous versus sequential lineup presentation. *Journal of Applied Psychology*, 70(3), 556–564. <https://doi.org/10.1037/0021-9010.70.3.556>
- Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods*, 49(2), 433–442. <https://doi.org/10.3758/s13428-016-0727-z>
- Loftus, G. R., & Harley, E. M. (2005). Why is it easier to identify someone close than far away? *Psychonomic Bulletin & Review*, 12(1), 43–65. <https://doi.org/10.3758/BF03196348>
- Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). Lawrence Erlbaum.
- Maniscalco, B., & Lau, H. (2014). Signal detection theory analysis of Type 1 and Type 2 data: Meta- d' , response-specific meta- d' , and the unequal variance SDT model. In S. M. Fleming & C. D. Frith (Eds.), *The cognitive neuroscience of metacognition* (pp. 25–66). Springer. https://doi.org/10.1007/978-3-642-45190-4_3
- Mansour, J. K., Beaudry, J. L., Groncki, R., & Nguyen, M. (2024). Eyewitness decision processes: A valid reflector variable. *Applied Cognitive Psychology*, 38(6), Article e70001. <https://doi.org/10.1002/acp.70001>
- Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. University of Minnesota Press. <https://doi.org/10.1037/11281-000>
- Metcalfe, J., & Finn, B. (2008). Evidence that judgments of learning are causally related to study choice. *Psychonomic Bulletin & Review*, 15(1), 174–179. <https://doi.org/10.3758/PBR.15.1.174>
- Mickes, L., Flowe, H. D., & Wixted, J. T. (2012). Receiver operating characteristic analysis of eyewitness memory: Comparing the diagnostic accuracy of simultaneous versus sequential lineups. *Journal of Experimental Psychology: Applied*, 18(4), 361–376. <https://doi.org/10.1037/a0030609>
- 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. <https://doi.org/10.1016/j.jarmac.2014.04.007>
- Molinaro, P. F., Charman, S. D., & Wylie, K. (2021). Pre-identification confidence is related to eyewitness lineup identification accuracy across heterogeneous encoding conditions. *Law and Human Behavior*, 45(6), 524–541. <https://doi.org/10.1037/lhb0000452>
- Moreland, M. B., & Clark, S. E. (2020). Absolute and relative decision processes in eyewitness identification. *Applied Cognitive Psychology*, 34(1), 142–156. <https://doi.org/10.1002/acp.3602>
- Nyman, T. J., Lampinen, J. M., Antfolk, J., Korkman, J., & Santtila, P. (2019). The distance threshold of reliable eyewitness identification. *Law and Human Behavior*, 43(6), 527–541. <https://doi.org/10.1037/lhb0000342>
- Odegard, T. N., & Lampinen, J. M. (2005). Recollection rejection: Gist cuing of verbatim memory. *Memory & Cognition*, 33(8), 1422–1430. <https://doi.org/10.3758/BF03193375>
- Palmer, M. A., Brewer, N., Weber, N., & Nagesh, A. (2013). The confidence–accuracy relationship for eyewitness identification decisions: Effects of exposure duration, retention interval, and divided attention. *Journal of Experimental Psychology: Applied*, 19(1), 55–71. <https://doi.org/10.1037/a0031602>
- Perfect, T. J. (2002). When does eyewitness confidence predict performance? In T. J. Perfect & B. L. Schwartz (Eds.), *Applied metacognition* (pp. 95–120). Cambridge University Press. <https://doi.org/10.1017/CBO9780511489976.006>
- Perfect, T. J., & Hollins, T. S. (1996). Predictive feeling of knowing judgements and postdictive confidence judgements in eyewitness memory and general knowledge. *Applied Cognitive Psychology*, 10(5), 371–382. [https://doi.org/10.1002/\(SICI\)1099-0720\(199610\)10:5<371::AID-ACP389>3.0.CO;2-O](https://doi.org/10.1002/(SICI)1099-0720(199610)10:5<371::AID-ACP389>3.0.CO;2-O)
- Perretta, S. F., & Dunning, D. (2001). *Testing the automatic versus process of elimination distinction in differentiating accurate from inaccurate eyewitnesses* [Unpublished manuscript]. Cornell University.
- Police Executive Research Forum. (2013). *A national survey of eyewitness identification procedures in law enforcement agencies*. https://www.policeforum.org/assets/docs/Free_Online_Documents/Eyewitness_Identification/a%20national%20survey%20of%20eyewitness%20identification%20procedures%20in%20law%20enforcement%20agencies%202013.pdf
- Price, H. L., Bruer, K. C., & Adkins, M. C. (2020). Using machine learning analyses to explore relations between eyewitness lineup looking behaviors and suspect guilt. *Law and Human Behavior*, 44(3), 223–237. <https://doi.org/10.1037/lhb0000364>
- Quigley-McBride, A., & Wells, G. L. (2023). Eyewitness confidence and decision time reflect identification accuracy in actual police lineups. *Law and Human Behavior*, 47(2), 333–347. <https://doi.org/10.1037/lhb0000518>
- R Core Team. (2024). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. <https://www.r-project.org>

- 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), Article 77. <https://doi.org/10.1186/1471-2105-12-77>
- Rotello, C. M., Macmillan, N. A., & Van Tassel, G. (2000). Recall-to-reject in recognition: Evidence from ROC curves. *Journal of Memory and Language*, 43(1), 67–88. <https://doi.org/10.1006/jmla.1999.2701>
- RStudio Team. (2020). *RStudio: Integrated development for R* [Computer software]. <https://www.rstudio.com/>
- Sauer, J. D., Brewer, N., & Weber, N. (2008). Multiple confidence estimates as indices of eyewitness memory. *Journal of Experimental Psychology: General*, 137(3), 528–547. <https://doi.org/10.1037/a0012712>
- Sauer, J. D., Brewer, N., & Wells, G. L. (2008). Is there a magical time boundary for diagnosing eyewitness identification accuracy in sequential line-ups? *Legal and Criminological Psychology*, 13(1), 123–135. <https://doi.org/10.1348/135532506X159203>
- Sauerland, M., Sagana, A., & Sporer, S. L. (2012). Assessing nonchoosers' eyewitness identification accuracy from photographic showups by using confidence and response times. *Law and Human Behavior*, 36(5), 394–403. <https://doi.org/10.1037/h0093926>
- Sauerland, M., & Sporer, S. L. (2007). Post-decision confidence, decision time, and self-reported decision processes as postdictors of identification accuracy. *Psychology, Crime & Law*, 13(6), 611–625. <https://doi.org/10.1080/10683160701264561>
- Sauerland, M., & Sporer, S. L. (2009). Fast and confident: Postdicting eyewitness identification accuracy in a field study. *Journal of Experimental Psychology: Applied*, 15(1), 46–62. <https://doi.org/10.1037/a0014560>
- Seale-Carlisle, T. M., Grabman, J. H., & Dodson, C. S. (2022). The language of accurate and inaccurate eyewitnesses. *Journal of Experimental Psychology: General*, 151(6), 1283–1305. <https://doi.org/10.1037/xge0001152>
- Shekhar, M., & Rahnev, D. (2024). How do humans give confidence? A comprehensive comparison of process models of perceptual metacognition. *Journal of Experimental Psychology: General*, 153(3), 656–688. <https://doi.org/10.1037/xge0001524>
- Smith, A. M., & Ayala, N. T. (2021). Do traditional lineups undermine the capacity for eyewitness memory to rule out innocent suspects? *Journal of Applied Research in Memory and Cognition*, 10(2), 215–220. <https://doi.org/10.1016/j.jarmac.2021.03.003>
- Smith, A. M., Ayala, N. T., & Ying, R. C. (2023). The rule out procedure: A signal-detection-informed approach to the collection of eyewitness identification evidence. *Psychology, Public Policy, and Law*, 29(1), 19–31. <https://doi.org/10.1037/law0000373>
- Smith, A. M., Smalarz, L., & Jalava, S. T. (2021). Measuring performance from eyewitness identification procedures. In A. M. Smith, M. Toglia, & J. M. Lampinen (Eds.), *Methods, measures, and theories in eyewitness identification tasks* (pp. 169–191). Taylor & Francis. <https://doi.org/10.4324/9781003138105-11>
- Smith, A. M., & Wells, G. L. (2023). Using reflector variables to determine whether the culprit is present in or absent from a police lineup. In K. Scherr & D. DeMatteo (Eds.), *The Oxford handbook of psychology and law* (pp. 464–479). Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780197649138.013.27>
- Smith, A. M., Yang, Y., & Wells, G. L. (2020). Distinguishing between investigator discriminability and eyewitness discriminability: A method for creating full receiver operating characteristic curves of lineup identification performance. *Perspectives on Psychological Science*, 15(3), 589–607. <https://doi.org/10.1177/1745691620902426>
- Smith, A. M., Ying, R. C., Ayala, N. T., & Goldstein, A. R. (2024). The untapped potential of lineups: Using eyewitness memory to rule out innocent suspects. *Psychology, Crime & Law*, 30(10), 1580–1590. <https://doi.org/10.1080/1068316X.2023.2242998>
- Smith, A. M., Ying, R. C., Goldstein, A. R., & Fitzgerald, R. J. (2024). Absolute-judgment models better predict eyewitness decision-making than do relative-judgment models. *Cognition*, 251, Article 105877. <https://doi.org/10.1016/j.cognition.2024.105877>
- Smith, S. M., Lindsay, R. C. L., & Pryke, S. (2000). Postdictors of eyewitness errors: Can false identifications be diagnosed? *Journal of Applied Psychology*, 85(4), 542–550. <https://doi.org/10.1037/0021-9010.85.4.542>
- Sporer, S. L. (1992). Post-dicting eyewitness accuracy: Confidence, decision-times and person descriptions of choosers and non-choosers. *European Journal of Social Psychology*, 22(2), 157–180. <https://doi.org/10.1002/ejsp.2420220205>
- Sporer, S. L. (1993). Eyewitness identification accuracy, confidence, and decision times in simultaneous and sequential lineups. *Journal of Applied Psychology*, 78(1), 22–33. <https://doi.org/10.1037/0021-9010.78.1.22>
- Sporer, S. L., Penrod, S., Read, D., & Cutler, B. (1995). Choosing, confidence, and accuracy: A meta-analysis of the confidence–accuracy relation in eyewitness identification studies. *Psychological Bulletin*, 118(3), 315–327. <https://doi.org/10.1037/0033-2909.118.3.315>
- Starns, J. J., Cohen, A. L., & Rotello, C. M. (2023). A complete method for assessing the effectiveness of eyewitness identification procedures: Expected information gain. *Psychological Review*, 130(3), 677–719. <https://doi.org/10.1037/rev0000332>
- 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(1), 99–139. <https://doi.org/10.1037/a0021650>
- Stebly, N. K., Dysart, J., Fulero, S., & Lindsay, R. C. (2001). Eyewitness accuracy rates in sequential and simultaneous lineup presentations: A meta-analytic comparison. *Law and Human Behavior*, 25(5), 459–473. <https://doi.org/10.1023/A:1012888715007>
- Stephens, R. G., Semmler, C., & Sauer, J. D. (2017). The effect of the proportion of mismatching trials and task orientation on the confidence–accuracy relationship in unfamiliar face matching. *Journal of Experimental Psychology: Applied*, 23(3), 336–353. <https://doi.org/10.1037/xap0000130>
- Strack, F., & Bless, H. (1994). Memory for nonoccurrences: Metacognitive and presuppositional strategies. *Journal of Memory and Language*, 33(2), 203–217. <https://doi.org/10.1006/jmla.1994.1010>
- Swets, J. A. (1973). The relative operating characteristic in psychology: A technique for isolating effects of response bias finds wide use in the study of perception and cognition. *Science*, 182(4116), 990–1000. <https://doi.org/10.1126/science.182.4116.990>
- Thompson, V. A., Prowse Turner, J. A., & Pennycook, G. (2011). Intuition, reason, and metacognition. *Cognitive Psychology*, 63(3), 107–140. <https://doi.org/10.1016/j.cogpsych.2011.06.001>
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267–288. <https://doi.org/10.1111/j.2517-6161.1996.tb02080.x>
- Tibshirani, R. J., & Wasserman, L. A. (2017). *Sparsity, the lasso, and friends*. <https://www.stat.cmu.edu/~ryantibs/statml/lectures/sparsity.pdf>
- Wagenaar, W. A., & van der Schrier, J. H. (1996). Face recognition as a function of distance and illumination: A practical tool for use in the courtroom. *Psychology, Crime & Law*, 2(4), 321–332. <https://doi.org/10.1080/10683169608409787>
- Weber, N., Brewer, N., Wells, G. L., Semmler, C., & Keast, A. (2004). Eyewitness identification accuracy and response latency: The unruly 10–12-second rule. *Journal of Experimental Psychology: Applied*, 10(3), 139–147. <https://doi.org/10.1037/1076-898X.10.3.139>
- Wells, G. L. (1984). The psychology of lineup identifications. *Journal of Applied Social Psychology*, 14(2), 89–103. <https://doi.org/10.1111/j.1559-1816.1984.tb02223.x>
- Wells, G. L. (1993). What do we know about eyewitness identification? *American Psychologist*, 48(5), 553–571. <https://doi.org/10.1037/0003-066X.48.5.553>

- Wells, G. L. (2020). Psychological science on eyewitness identification and its impact on police practices and policies. *American Psychologist*, 75(9), 1316–1329. <https://doi.org/10.1037/amp0000749>
- Wells, G. L., Kovera, M. B., Douglass, A. B., Brewer, N., Meissner, C. A., & Wixted, J. T. (2020). Policy and procedure recommendations for the collection and preservation of eyewitness identification evidence. *Law and Human Behavior*, 44(1), 3–36. <https://doi.org/10.1037/lhb0000359>
- Wells, G. L., & Luus, C. E. (1990). Police lineups as experiments: Social methodology as a framework for properly conducted lineups. *Personality and Social Psychology Bulletin*, 16(1), 106–117. <https://doi.org/10.1177/0146167290161008>
- 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(6), 603–647. <https://doi.org/10.1023/A:1025750605807>
- Wells, G. L., Steblay, N. K., & Dysart, J. E. (2015). Double-blind photo lineups using actual eyewitnesses: An experimental test of a sequential versus simultaneous lineup procedure. *Law and Human Behavior*, 39(1), 1–14. <https://doi.org/10.1037/lhb0000096>
- Wells, G. L., & Windschitl, P. D. (1999). Stimulus sampling and social psychological experimentation. *Personality and Social Psychology Bulletin*, 25(9), 1115–1125. <https://doi.org/10.1177/01461672992512005>
- Wells, G. L., Yang, Y., & Smalarz, L. (2015). Eyewitness identification: Bayesian information gain, base-rate effect equivalency curves, and reasonable suspicion. *Law and Human Behavior*, 39(2), 99–122. <https://doi.org/10.1037/lhb0000125>
- Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., ... Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43), Article 1686. <https://doi.org/10.21105/joss.01686>
- Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2023). *dplyr: A grammar of data manipulation* (R package Version 1.1.4) [Computer software]. <https://dplyr.tidyverse.org>
- Wixted, J. T., & Mickes, L. (2014). A signal-detection-based diagnostic-feature-detection model of eyewitness identification. *Psychological Review*, 121(2), 262–276. <https://doi.org/10.1037/a0035940>
- Wixted, J. T., Mickes, L., Dunn, J. C., Clark, S. E., & Wells, W. (2016). Estimating the reliability of eyewitness identifications from police lineups. *Proceedings of the National Academy of Sciences of the United States of America*, 113(2), 304–309. <https://doi.org/10.1073/pnas.1516814112>
- Wixted, J. T., & Wells, G. L. (2017). The relationship between eyewitness confidence and identification accuracy: A new synthesis. *Psychological Science in the Public Interest*, 18(1), 10–65. <https://doi.org/10.1177/1529100616686966>
- Yilmaz, A. S., Lebensfeld, T. C., & Wilson, B. M. (2022). The reveal procedure: A way to enhance evidence of innocence from police lineups. *Law and Human Behavior*, 46(2), 164–173. <https://doi.org/10.1037/lhb0000478>

Received August 28, 2024

Revision received October 30, 2024

Accepted November 11, 2024 ■