UNIVERSITY OF CALIFORNIA SAN DIEGO

The Psychology of Lineup Rejections in Eyewitness Identification

A dissertation submitted in partial satisfaction of the
requirements for the degree of Doctor of Philosophy

in

Experimental Psychology with Specialization in Anthropogeny

by

Anne Sheyda Yilmaz

Committee in charge:

Professor John Wixted, Chair
Professor John Serences
Professor Timothy Brady
Professor Uma Karmarkar

2024

The Dissertation of Anne Sheyda Yilmaz is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2024

DEDICATION

To Dr. John Wixted who took a chance on a student who knew nothing: Thank you for believing in me. (And then, thank you for putting up with me for many more years than you bargained for.) You're a fantastic mentor, and I owe much of my academic success to your willingness to invest in me as a person and as a researcher.

To Dr. Pascal Gagneux, who is a whirlwind of knowledge and care: Thank you for inspiring me and for assuring me early on that naivete is nothing to be insecure about—it is simply the marker that you have the privilege to learn more.

To the Center for Academic Research and Training in Anthropogeny (CARTA), who funded me on fellowship for two years and took me across the world for research: Thank you for cultivating an environment of curiosity and collaboration across domains, for an appreciation of field work, and for the life-altering experience of meeting the Hadza.

To all my past teachers and professors who served as touchstones on my academic journey: Thank you for guiding me.

To my mom and to my friends: Thank you for supporting me in my absence from the "real world," despite not fully knowing what the heck one does in a doctoral program or why it was taking so long. Thank you for your relentless patience as my life has, in many ways, been on hold for the last few years.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

Chapter 1 is a reprint of the abstract of the broader material as it appears in: Yilmaz, A.S., Lebensfeld, T., & Wilson, B.M. (2022). Enhancing evidence of innocence from police lineups. *Law and Human Behavior, 46*(2), 164-173. The dissertation author was the primary researcher and author of this paper. Permission to use the published abstract as it appears was granted by the American Psychological Association, publisher of *Law and Human Behavior.* All co-authors (Taylor Lebensfeld and Brent Wilson) and the dissertation committee chair (Professor John Wixted) have given permission to use this work in fulfillment of my dissertation requirements.

Chapter 2, in full, is a reprint of the material as it appears in: Yilmaz, A.S**.** & Wixted, J.T. (2024). What latent variable underlies confidence in lineup rejections? *Journal of Memory and Language*, *135,* 104493. The dissertation author was the primary researcher and author of this paper. A full reprint of this material within a dissertation is allowed without permission by Elsevier, the publisher of *Journal of Memory and Language*, as long as the dissertation author is also the author of the original paper and the dissertation is not published commercially. Professor John Wixted, who is both a co-author and the dissertation committee chair, has given permission to use this work in fulfillment of my dissertation requirements.

Chapter 3, in full, is a reprint of the material as it appears in: Yilmaz, A.S., Wang, X., & Wixted, J.T. (2024). Response bias modulates the confidence-accuracy relationship for both positive IDs and lineup rejections in a simultaneous lineup task. *Applied Cognitive Psychology*, 38(2), e4196. The dissertation author was the primary researcher

and author of this material. Permission to use the material as it appears was granted by John Wiley and Sons, the publisher of *Applied Cognitive Psychology*. All co-authors (Xiaoqing Wang and Professor John Wixted) and the dissertation committee chair (Professor John Wixted) have given permission to use this work in fulfillment of my dissertation requirements.

VITA

2016	Bachelor of Science, University of Oregon and the Robert Donald Clark Honors College
2017	Research Assistant, University of California San Diego
2018-2024	Teaching Assistant, University of California San Diego
2020	Master of Arts, University of California San Diego
2023-2024	Associate-in-Lieu, University of California San Diego
2024	Doctor of Philosophy, University of California San Diego

PUBLICATIONS

Yilmaz, A.S., Wilson, B.M., & Wixted, J.T. (2024). A Rate-them-all lineup procedure increases information and reduces discriminability. *Journal of Experimental Psychology: Applied*. Advance online publication. https://doi.org/10.1037/xap0000524

Yilmaz, A.S., Wang, X., & Wixted, J.T. (2024). Response bias modulates the confidence-accuracy relationship for both positive IDs and lineup rejections in a simultaneous lineup task. *Applied Cognitive Psychology*, 38(2), e4196. https://doi.org/10.1002/acp.4196

Yilmaz, A.S**.** & Wixted, J.T. (2024). What latent variable underlies confidence in lineup rejections? *Journal of Memory and Language*, *135,* 104493. https://doi.org/10.1016/j.jml.2023.104493

Yilmaz, A.S., Laney, C., Loftus, E.F., Spielman, R.M., Stangor, C., & Walinga, J. (2023). Memory in context. In C. Pilegard (Ed.), *Cognitive Foundations* (2nd ed.).

Yilmaz, A.S., Lebensfeld, T., & Wilson, B.M. (2022). Enhancing evidence of innocence from police lineups. *Law and Human Behavior, 46*(2), 164-173. https://doi.org/10.1037/lhb0000478

Yilmaz, A.S. (2016). Eyewitness memory: How stress and situational factors affect eyewitness recall (Undergraduate honors thesis). For the Robert D. Clark Honors College and University of Oregon Department of Psychology, Eugene, USA.

FIELDS OF STUDY

Major Field: Experimental Psychology

Studies in Cognitive Psychology and Anthropogeny
Professors John Wixted and Pascal Gagneux

ABSTRACT OF THE DISSERTATION

The Psychology of Lineup Rejections in Eyewitness Identification

by

Anne Sheyda Yilmaz

Doctor of Philosophy in Experimental Psychology
with Specialization in Anthropogeny

University of California San Diego, 2024

Professor John Wixted, Chair

In recent years, the field has found that the confidence-accuracy relationship for positive identifications (ID) made from a police lineup is often strong while the relationship for lineup rejections is typically much weaker. The reason for this asymmetry remains unclear. Here, we report results from signal-detection-based simulations and

models, as well as from mock-crime lineup experiments, to help explain why this is often observed. When a face is identified from a photo lineup, the selected face is presumably the one that generates the strongest memory signal, with confidence presumably being determined by the strength of the signal associated with that face. When a lineup is rejected, the entire set of faces is collectively rejected due to no face generating a sufficiently strong memory signal to be identified. One theory is that confidence for rejections is determined by the average strength of the memory signals instead of the singular memory signal generated by the most familiar face. Averaging could wash out what would otherwise be a strong confidence-accuracy relationship. Chapter 1 investigated whether changing the lineup task such that participants reject only a singular face instead of a set of faces will strengthen the confidence-accuracy relationship for rejections. We found support for this hypothesis. Chapter 2 used multiple data sets for an in-depth modeling paper investigating whether the averaging of signals is the basis for confidence in a rejection. Our model-fitting analysis found that confidence in a lineup rejection is not based on the average signal and is instead based on the most familiar face, just as is the case for positive IDs. Chapter 3 investigated whether response bias, not averaging, may be a determinant of the strength of the confidence-accuracy relationship. Inducing a more conservative response bias should theoretically weaken the relationship for positive IDs while strengthening it for lineup rejections because a conservative criterion shift increases the range of possible memory signals associated with that decision. Our results support this prediction, showing that the degree of range restriction directly corresponds to the strength of the confidence-accuracy relationship for lineup rejections.

INTRODUCTION

In the last 10 years, the field of eyewitness identification experienced a change in the way eyewitness decisions are understood. This change was brought about by bringing signal detection theory (SDT) methods long used in cognitive psychology to lineup decision-making (Wixted & Mickes, 2014). Although initially controversial, the SDT methodology of receiver operating characteristic (ROC) analysis was endorsed by the National Academies of Sciences in 2014 (and again by co-chairs Albright & Rakoff of that committee in 2022), which solidified its use in the eyewitness domain.

The most common way of testing an eyewitness's memory for a suspect is the simultaneous photo lineup (Police Executive Research Forum, 2013). Ideally, the lineup would consist of one photo of the suspect (i.e., the person the police believe may have committed the crime) and five or more similar looking "fillers" who are known to be innocent. The witness can either make a positive identification (an "ID"; landing on the suspect or on one of the fillers) or reject the lineup (declaring that none of the faces in the lineup matches the witness's memory of the perpetrator). Regardless of the decision, the witness is often asked to provide a confidence rating as well.

Before the application of SDT methodology, the field believed that confidence and accuracy were not strongly related to one another in the case of positive identifications (i.e., when a person is selected from a lineup as being guilty). However, with analytical guidance from SDT, the field came to understand that the opposite is true (Gronlund et al., 2014; Wixted et al., 2015; Wixted & Wells, 2017). Confidence is strongly predictive of accuracy for positive IDs on the first lineup test in the sense that high-confidence identifications are highly accurate and low-confidence identifications are often inaccurate.

The discovery that positive IDs can have high information value led the field to seek a better understanding of the decision variable that witnesses use to make a positive ID and to determine confidence. In this regard, researchers have proposed three different models of decision-making to explain how identifications in lineups worked, specifically, the Independent Observations model, the Integration model, and the Ensemble model (Wixted et al., 2018). According to the Independent Observations model, the witness first singles out the face that most strongly matches their memory of the perpetrator (often described as the "MAX" face) without regard for the memory signals generated by the other faces in the lineup. If the memory signal associated with that face exceeds the criterion for declaring a face as being sufficiently familiar, then an identification is made. If not, both the MAX face and the lineup as a whole are rejected. If a positive ID is made, confidence is determined by the strength of the memory signal in relation to additional confidence criteria. If a memory signal strength is high enough to exceed the "high-confidence" criterion, then the identification is made with high confidence. If the memory signal strength falls just shy of that high-confidence criterion, then the identification is made with medium confidence, and so on.

The Integration model is different from the Independent Observations model in that the MAX face is not the initial focus of attention (Wixted et al., 2018). Instead, the sum of memory signal strengths for all of the faces in the lineup is considered. If that sum exceeds the decision criterion, then the MAX face is identified. A face is not selected if the sum of memory signal strengths does not exceed the decision criterion. Confidence in the case of a positive ID is determined in a fashion similar to that of the Independent Observations

model: Confidence corresponds to the highest confidence criterion that the summed memory signal surpasses.

In the Ensemble model, the average memory signal for all of the faces in the lineup is first considered (Wixted et al., 2018). Then, a difference score between the MAX face and the lineup average is computed. The MAX face is selected as the perpetrator if the difference between the MAX and the average (i.e., if the ensemble decision variable) exceeds the decision criterion. The more that difference score exceeds the criterion, the higher the confidence. In essence, confidence is higher the more the memory signal for the MAX face stands out from the crowd of memory signals generated by the faces in the lineup. According to one recent review of the literature, the Ensemble model appears to be the most viable model of the decision variable used by witnesses in the case of positive IDs in a police lineup (Wixted et al., 2018).

Things are much less clear when it comes to the decision variable witnesses use to determine their confidence when they reject a lineup. For example, in contrast to the strong relationship for positive IDs, the relationship between confidence and accuracy for lineup rejections is often (but not always) negligible (Brewer & Wells, 2006; Arndorfer & Charman, 2022). Thus, a high-confidence rejection is not necessarily indicative of high accuracy like it is in the case of a positive ID. One explanation for this difference might be that, in a lineup rejection, confidence is not tied to an individual face like it is for a positive ID. Instead, participants might reject the lineup as a whole.

If a lineup is rejected because, for example, the ensemble decision variable (MAX – mean) for each of the lineup faces fails to exceed the decision criterion, what decision variable determines confidence? Confidence might still be based on the MAX minus mean

decision variable, though a reasonably strong confidence-accuracy relationship would be expected in that case. It is possible that participants instead average (AVG) the memory signals of all of the faces in the lineup (similar to what is done in the Ensemble model, but without the selection of a MAX face) and the confidence rating is based upon said average. Rudimentary signal detection simulations recently reported by Yilmaz et al. (2022) indicate that the AVG rule can reduce the confidence-accuracy relationship that would be present if faces were individuated in that decision instead. Also, it is possible that another decision variable entirely—something other than the AVG or the MAX—is being used. Regardless, it seems reasonable to assume that the differences in confidence-accuracy relationship for positive IDs vs. lineup rejections may be caused by the use of a different decision variable. In fact, this explanation was first proposed long ago by Brewer and Wells (2006).

Another possible explanation for the weak confidence-accuracy relationship for lineup rejections focuses on response bias instead of the decision variable. Theoretically, if participants had a liberal response criterion, the range of memory signal strengths associated with a positive ID would increase, and the increased range would make it easier to detect a confidence-accuracy relationship. In turn, the same liberal response criterion would commensurately shrink the range of memory signal strengths corresponding to a lineup rejection, and the decreased range would make it harder to detect a relationship. A conservative response criterion, on the other hand, would theoretically cause the inverse to be true. It may be that the studies within the literature primarily induce more-liberal responding, which could be why we often see a strong confidence-accuracy relationship for positive IDs but not for rejections. Studies which show a stronger-than-typical relationship for rejections may have simply induced more conservative responding.

Broadly, the goal of this dissertation is to shed light on why the diagnostic value of confidence is lower in the cases of lineup rejections, and to work to improve it. The specific questions asked to work toward this goal are:

1) Will a change in lineup procedure in the case of rejections—such that the confidence value is tied to the singular suspect's face instead of the set of faces—successfully increase the strength of the confidence-accuracy relationship for lineup rejections?

2) Can we determine whether the MAX or AVG decision rule has greater support by comparing goodness of fit using the Independent Observations model and Ensemble model across multiple data sets?

3) Is the weak confidence-accuracy relationship often observed for lineup rejections a byproduct of response bias affecting the range of memory signals associated with positive IDs and rejections?

The dissertation will consist of three published studies. Each study will directly address one of the three questions listed above.

Chapter 1 aimed to answer the first question above, *"Will a change in lineup procedure in the case of rejections—such that the confidence value is tied to the singular suspect's face instead of the set of faces—successfully increase the strength of the confidence-accuracy relationship for lineup rejections?"* In Chapter 1, we compared a standard simultaneous lineup to the Reveal procedure. The Reveal procedure was identical to the standard simultaneous lineup except in the case of lineup rejections. Typically, as is the case with the standard lineup, when a lineup was rejected, participants immediately

gave their confidence rating. For the Reveal procedure, after a lineup rejection occurred, *but before confidence was gathered*, the suspect popped up on the screen asking for participants to state their confidence that the suspect is NOT the perpetrator. Participants were *not* allowed to change their mind and say that the suspect on the screen matched their memory after all (indicating them to be the perpetrator from the mock crime video). They had to stay with the rejection decision. This design allows for lineup administrators to gather information directly about their suspect even in the case of a rejection and in a manner that would not imperil an innocent suspect before the witness made the decision. We analyzed the confidence-accuracy relationship using CAC analysis and found that the Reveal procedure increased both the accuracy and the frequency of high-confidence rejections compared to the standard simultaneous lineup procedure. Furthermore, this increase in high-confidence rejections occurred in cases in which the suspect was innocent, not guilty.

Chapter 2 of this dissertation aimed to answer the second question above, *"Can we determine whether the MAX or AVG decision rule has greater support by comparing goodness of fit using the Independent Observations model and Ensemble model across multiple data sets?"* In Chapter 2, we modified two existing signal-detection-based models (e.g., the Independent Observations model and Ensemble model) to determine whether the MAX, not the AVG, memory signal was the basis for witness confidence in the lineup rejections. Without modifications, Independent Observations model assumes that the MAX signal—which is considered independent of the memory signals generated by the other faces—is the basis for confidence for both positive identifications and lineup rejections. If the MAX signal passes the decision criterion, then a face is selected, and the distance

between the signal and the decision criterion determines confidence. The further away the signal is from the criterion, the greater the confidence in the selection. When the MAX signal fails to pass the criterion, a lineup rejection occurs and, similarly, the further away the MAX signal is from the criterion (just in the opposite direction), the higher the level of confidence in the rejection.

The Ensemble model also assumes a MAX rule as the basis of confidence for positive identifications and rejections. The difference between the Independent Observation model and the Ensemble model is that, as indicated above, the Independent Observations model considers the MAX signal independent of the other signals while the Ensemble model considers a transformed MAX signal. For the Ensemble model, a difference score between the MAX face and the mean of all of the faces is computed (MAX – mean). This allows the strength of the MAX face to be considered relative to how much that MAX face stands out from the other faces in the ensemble. If that transformed MAX signal exceeds the decision criterion, a selection is made and the distance between the transformed memory signal and the decision criterion determines confidence. A lineup rejection occurs when the transformed MAX signal fails to pass the decision criterion. The further away the transformed MAX signal is from the criterion in the direction of the rejection, the greater the confidence associated with the rejection decision.

We modified the Independent Observations and Ensemble models to assume that the AVG memory was the decision variable for confidence in the case a rejection. The AVG decision variable used for lineup rejections was the same for both models. To do this, we assumed that the AVG of all of the memory signals generated by the faces, not the MAX signal or transformed MAX signal, was the basis for confidence. Specifically, if the AVG

signal was closer to the decision criterion (but still failing to pass it), a rejection would occur with lower levels of confidence. If the AVG signal was further away from the decision criterion in the negative direction, the rejection would be made with a higher level of confidence. The derivation of the AVG likelihood function followed steps similar to those described in Wixted et al. (2018), and we used simulated data to verify whether the modified likelihood function was correct. Based on the goodness of fit of these AVG models, we determined that the MAX, not the AVG, memory signal is the basis for witness confidence in the lineup rejections. Although this does not rule out the AVG model completely, much stronger support was found for the MAX models compared to the AVG models across datasets.

Chapter 3 aimed to answer the last question of the dissertation, *"Is the weak confidence-accuracy relationship often observed for lineup rejections a byproduct of response bias affecting the range of memory signals associated with positive IDs and rejections?"* In this study, participants viewed a simultaneous lineup after a mock crime video. This lineup was a forced-choice procedure in which they had to choose someone in the lineup as being the perpetrator from a mock crime video. Participants provided their confidence level using a -100 to +100 scale, with the negative values corresponding to the participants' belief that the person they were forced to select was innocent. This design allowed us to effectively manipulate participants' willingness to respond (i.e., response bias) on a monotonic scale. More specifically, we manually varied the position of the decision criterion by setting it to be highly conservative (+80), highly liberal (-80), or to other less extreme biases (e.g., -50, 0, 50+). This allowed us to view the confidence-accuracy relationship for positive IDs and lineup rejections for the same procedure as

function of response bias. As predicted, we found that a conservative response criterion led to a flatter-but-highly accurate CAC for positive IDs, and a steeper CAC for lineup rejections. In turn, a liberal decision criterion led to a steeper CAC for positive IDs, and a flatter CAC for lineup rejections.

Much has been learned over the last decade of eyewitness research due to the influence of signal detection theory, but much of that work has focused on what occurs when a witness makes a selection from a lineup. Comparatively, there is a lack of understanding of what happens when a witness rejects a lineup, and working toward further understanding the information value of rejection decisions could provide important insight to real-world cases. For example, in the DNA exoneration cases reported by the Innocence Project, witnesses who confidently misidentified an innocent defendant at trial often initially rejected the lineup (or picked a filler photo) on their first test of memory (Garrett, 2011). Additionally, of the 208 eyewitness cases in the National Registry of Exonerations that contained information about the initial identification, 190 cases had witnesses who rejected the initial lineup containing the innocent suspect despite that witness later misidentifying the same innocent suspect (Yilmaz et al., 2024a).

AMERICAN PSYCHOLOGICAL ASSOCIATION
American Psychology-Law Society
Law and Human Behavior

ISSN: 0147-7307
2022, Vol. 46, No. 2, 164–173
https://doi.org/10.1037/lhb0000478

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

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

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

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

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

Supplemental materials: https://doi.org/10.1037/lhb0000478.supp

ACKNOWLEDGEMENTS

Chapter 1 is a reprint of the abstract of the broader dissertation work as it appears in: Yilmaz, A.S., Lebensfeld, T., & Wilson, B.M. (2022). Enhancing evidence of innocence from police lineups. *Law and Human Behavior, 46*(2), 164-173. The dissertation author was the primary researcher and author of this paper. Permission to use the material as it appears was granted by the American Psychological Association, publisher of *Law and Human Behavior.* All co-authors (Taylor Lebensfeld and Brent Wilson) and the dissertation committee chair (Professor John Wixted) have given permission to use this work in fulfillment of my dissertation requirements.

Journal of Memory and Language 135 (2024) 104493

Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

What latent variable underlies confidence in lineup rejections?

Anne S. Yilmaz, John T. Wixted [*,1]

University of California, San Diego, USA

A B S T R A C T

When a face is positively identified from a multi-person photo lineup, it is presumably the face that generates the strongest memory signal. In addition, confidence in a positive identification is presumably determined by the strength of the memory signal associated with that face. However, when no face generates a strong enough memory signal to be identified, the entire set of faces in the lineup is collectively rejected. What latent variable underlies confidence in a lineup rejection? One possibility is that the face that generates the strongest memory signal still determines confidence (i.e., the weaker that memory signal is, the more confidently the lineup is rejected). Another possibility is that confidence in a lineup rejection is determined by the average strength of the memory signals generated by the faces in the lineup (i.e., the weaker that average memory signal is, the more confidently the lineup is rejected). The reliance on an average signal has been proposed as a possible explanation for why the confidence-accuracy for lineup rejections tends to be weak. Here, we modified two existing signal-detection-based lineup models (the Independent Observations model and the Ensemble model) and fit them to multiple lineup datasets to investigate which decision variable underlies confidence in lineup rejections. Both models agree that confidence in a lineup rejection is based on the strongest memory signal in the lineup, not on the average signal. These model fits also revealed for the first time that the memory signals in a lineup are correlated, as they theoretically should be.

Introduction

A theoretically interesting issue in the domain of recognition memory concerns the *decision variable* that participants use to decide whether an item was previously encountered. In a standard old/new recognition procedure, the decision variable is simply the memory signal generated by the singular item presented on a given test trial. The nature of this memory signal can be conceptualized in terms of recollection vs. familiarity, item vs. associative information, or verbatim vs. gist memory—but however it is conceptualized, the stronger that memory signal is, the more likely the test item is to be declared "old" and the higher the participant's confidence will be.

When more than one item is presented on a given test trial, other decision variables become possible. In a standard two-alternative forced-choice (2-AFC) procedure, for example, the item chosen on a given trial is presumably the one that generates the stronger memory signal. However, the participant's confidence in that choice could be based either on the strength of the winning item's memory signal considered in isolation (i.e., without regard for the strength of the losing item), or it could instead be based on the difference in memory strength associated with the two test items, in which case confidence would be higher the more the strength of the winning item exceeds that of the losing item. Ignoring the strength of the losing item is suboptimal in the sense that it leaves useful information on the table, but the results of a several recent studies have suggested that participants do just that (e.g., Hanczakowskia, Butowska, Beaman, Jones, Zawadzka, 2021; Jou, Flores, Cortes, & Leka, 2016; Miyoshi, Kuwahara, & Kawaguchi, 2018; Zawadzka, Higham, & Hanczakowski, 2017).

Similar theoretical issues arise when more items are presented on a test trial, such as in the case of a police photo lineup. A typical photo lineup consists of six or more faces that are arranged in one of two possible configurations. A *target-present* lineup consists of one previously seen "old" face (i.e., the target) surrounded by five or more new "fillers" (i.e., lures) that are drawn from a pool of photos all of which are matched to the target on basic characteristics like race, gender, hairstyle, and approximate age. A *target-absent* lineup is similar except that the target is replaced by another filler to serve as the "innocent suspect." An innocent suspect in an actual police lineup is special from the perspective of the police (being the only person in the lineup suspected of having committed the crime), but from the perspective of the witness, the innocent suspect is not special and is functionally just another filler (i.e., an innocent person who matches the other lineup members with respect to general physical characteristics). When presented with a lineup, participants can choose one of the faces as having been seen before or they can reject the lineup by indicating that the target is not present.

As in 2-AFC, if a face is chosen from a lineup, it is presumably the one that generates the strongest (MAX) memory signal. However, once

* Corresponding author at: Department of Psychology, University of California, San Diego, La Jolla, CA 92093, USA.
E-mail address: jwixted@ucsd.edu (J.T. Wixted).

[1] Supported by a grant from the UCSD Yankelovich Center and in part by the Center of Academic Research and Training in Anthropogeny (CARTA) Fellowship.

https://doi.org/10.1016/j.jml.2023.104493
Received 5 April 2023; Received in revised form 26 October 2023; Accepted 4 December 2023
Available online 28 December 2023

again, confidence in a positive identification might be based solely on the absolute strength of the memory signal associated with the chosen face (without regard for the strength of the other faces in the lineup) or it might instead be based on a difference score. A signal detection model known as the Independent Observations model assumes that confidence in a positive identification from a lineup is based on its absolute memory signal (Wixted et al., 2018). An alternative signal detection model known as the Ensemble model assumes that confidence in a positive identification from a lineup is instead based on a difference score. According to this model, confidence in a positive ID is based on the MAX signal minus the mean memory strength signal across all faces in the lineup. In that case, confidence would be high not merely when the MAX signal is strong (as is true of the Independent Observations model) but only when its high strength stands out sufficiently from the "crowd" of memory signals in the lineup (Akan et al., 2021; Wixted et al., 2018).

The research reported here does not address the absolute vs. relative issue for positive IDs but instead focuses on the largely unexplored decision variable that underlies confidence for negative IDs (i.e., for lineup rejections). Critically, unlike in the case of positive IDs, no face is selected when a lineup is rejected. In that case, is confidence still determined by the memory signal associated with the unchosen MAX face (either its absolute memory strength or its memory strength relative to the other faces in the lineup)? Or is it instead based on a collective memory signal, such as the average (AVG) of the memory signal generated by all the faces in a lineup?

It seems fair to say that the default view is that the confidence in lineup rejections is based on the MAX signal, just as is true of confidence in positive identifications (e.g., Akan et al., 2021). However, picking up on an idea suggested by Brewer and Wells (2006) and Lindsay et al. (2013), Yilmaz et al. (2022) hypothesized that confidence in lineup rejections might be determined by the average memory signal. The rationale for deviating from the default perspective was based on the empirical observation that the confidence-accuracy relationship for lineup rejections, unlike the confidence-accuracy for positive IDs, is often weak (e.g., Brewer & Wells, 2006) and is sometimes completely flat (e.g., Dodson & Dobolyi, 2016). One possible reason for that asymmetry is that a different decision variable is used for positive vs. negative IDs. It seems plausible that a different decision variable might be used because, for positive IDs, confidence is provided in relation to a single face (i.e., the MAX face), whereas for negative IDs (i.e., lineup rejections), confidence is provided to the *set* of rejected faces. Here, using a model-fitting approach, we investigate whether the MAX memory signal or the AVG memory signal underlies confidence in lineup rejections.

The primary goal of our model-fitting approach is to rule out the least viable model, leaving the winning model as a viable candidate. As noted by Roberts and Pashler (2000), the mere fact that a model provides a better fit cannot be assumed to validate that model. However, Wixted et al. (2018) argued that a model that provides a qualitatively poor fit relative to other models can be reasonably rejected. For example, for the fits reported by Wixted et al. (2018), the Integration model (according to which the decision variable is based on the sum of the memory signals associated with the individual faces in the lineup) provided a far worse fit to the data than the Independent Observations and Ensemble models. On those grounds, the Integration model was ruled out as a viable candidate. Our goal here is to determine if, for lineup rejections, the assumption of a MAX decision variable similarly provides a qualitatively worse fit to the data than a model based on an AVG decision variable, perhaps helping to explain the weak confidence-accuracy relationship when the witness decides that the perpetrator is not in the lineup.

To investigate this issue, we (1) modified both the Independent Observations model and the Ensemble model to use either a MAX decision variable or an AVG decision variable to determine confidence in lineup rejections (yielding two versions of each model) and then (2) fit those models to empirical lineup data to determine which better characterizes the results. According to the MAX version of each model, the weaker the (absolute or relative) signal associated with the MAX face is, the more confidently the lineup is rejected. According to the AVG version, the weaker the average signal associated with the set of faces in the lineup is, the more confidently the lineup is rejected.

Because the Independent Observations and Ensemble models used in prior research already assume that the MAX face determines confidence for positive IDs, extending that assumption to confidence in negative IDs required only minor changes. By contrast, modifying the two models to allow for the possibility of an AVG decision variable for lineup rejections was more involved because it required modifying the likelihood functions for positive IDs derived by Wixted et al. (2018). The next section describes how the Independent Observations model and the Ensemble model conceptualize confidence in positive IDs and then provides an overview of how their likelihood functions were modified to allow for the possibility that an AVG memory signal is used for confidence in lineup rejections (with the mathematical details presented in the Appendix).

Signal detection models of lineup memory

Basic assumptions

Fig. 1 illustrates a standard signal detection representation of the memory signals generated by faces in target-present and target-absent lineups. In a target-present lineup (top panel of Fig. 1), the raw memory-match signal for the guilty suspect (i.e., the degree to which the

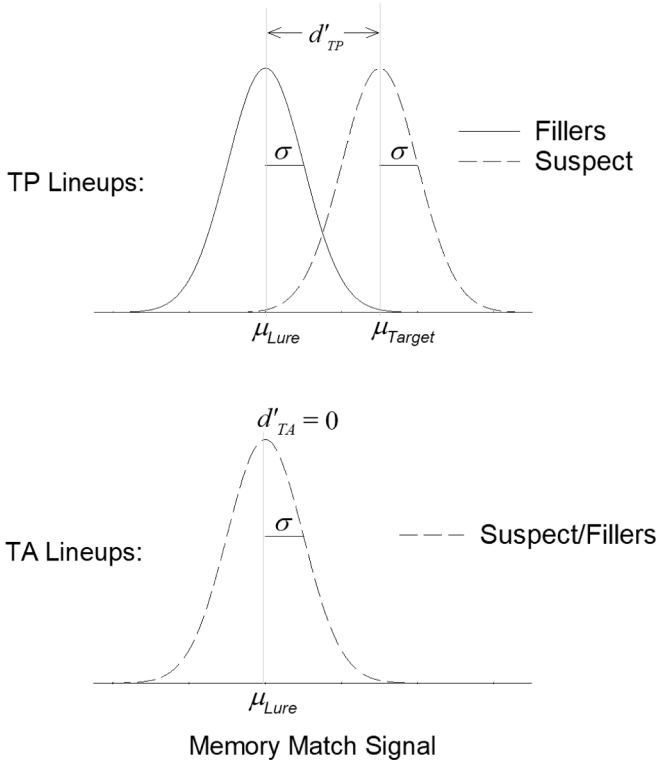

Fig. 1. Memory-match signals in target-present (TP) and target-absent (TA) lineups. μ_{Target} represents the mean of the guilty suspect distribution (the guilty suspect is the previously seen target). For the simplest case in which a single pool of fillers is used for all fillers and innocent suspects, the mean of the distribution of memory-match signals is μ_{Lure}, which can be set to zero for convenience. In target-present lineups, d'_{TP} is the difference between the mean of the guilty suspect (target) distribution and the lure distribution in standard deviation units. That is, $d'_{TP} = \frac{\mu_{Target} - \mu_{Lure}}{\sigma}$ for the uncorrelated case. Similarly, for target-absent lineups, d'_{TA} is the standardized difference between the innocent suspect distribution and the lure distribution. Because the innocent suspect is simply another face drawn from the pool of fillers, $d'_{TA} = 0$.

face of the guilty suspect in the lineup matches the face of the perpetrator in memory) is drawn from a distribution with a relatively high mean, whereas the memory-match signals for the fillers are drawn from a distribution with a lower mean. By contrast, in a target-absent lineup, the innocent suspect is effectively just another filler. Thus, the memory-strength distributions for the innocent suspect and the TA fillers are one and the same (bottom panel of Fig. 1).

The memory signals generated by the suspect and fillers in a lineup are likely to be positively correlated because the faces are not chosen randomly. Instead, to ensure a fair lineup, they are chosen because they share basic physical features of the perpetrator that are likely stored in the witness's memory, such as race, gender, age, etc. (Wells et al., 1998, Wells et al., 2020). In actual police investigations, witnesses often describe these features to the police, and a longstanding recommendation is that photos should be included in the lineup only if they match the witness's description of the perpetrator (Wells et al., 1993). The shared features are what give rise to correlated memory signals. For example, if an impoverished memory of the perpetrator was formed at the time the crime was witnessed, the shared features will not generate a strong memory-match signal, and this will be true of all the faces in the lineup. If a rich memory of the perpetrator was formed instead, the shared features will generate a strong memory-match signal, and, again, this will be true of all the faces in the lineup. Thus, the fact that features that are shared across faces in a lineup give rise to correlated memory signals is by design. This is an important issue that we return to later, but we set it aside for the moment to simplify the discussion of the likelihood functions for the competing models of interest here.

The distributions of raw memory signals shown in Fig. 1 serve as the general foundation of any signal detection model of recognition memory tested using a standard lineup. Specific models are created by specifying how those memory signals are used to make recognition memory decisions. The Independent Observations model and Ensemble model make different assumptions about how these memory signals are evaluated in relation to decision criteria to (1) make a decision about whether a face in the lineup is the previously seen perpetrator and (2) rate confidence when a face is identified.

Modeling positive IDs

The Independent Observations model assumes that positive IDs are based on the raw strength of the memory signals depicted in Fig. 1. Thus, according to this model, the overall decision criterion for making a positive ID and the additional criteria for rating confidence are superimposed on the distribution of raw memory-match signals shown in Fig. 1, as illustrated in Fig. 2. In Fig. 2, the upper and lower panels shown in Fig. 1 have been collapsed into a single panel because the distribution of memory signals for fillers in both target-present and target-absent lineups and for innocent suspects in target-absent lineups is the same (i.e., they are all faces drawn from the same pool of fillers).

The Independent Observations model assumes that the decision is based on the face in the lineup that generates the strongest memory-match signal (the MAX face), regardless of the memory-strength signals generated by the other faces. In other words, the decision is independent of the signals associated with those other faces. No face other than the MAX face has any bearing on the decision. If the memory signal of the MAX face in the lineup exceeds an overall decision criterion (c_3), then that face will be identified regardless of whether the memory signals generated by other faces in the lineup also happen to exceed the decision criterion (Macmillan & Creelman, 2005; Wixted et al., 2018). The stronger the memory signal generated by the MAX face is (e.g., if it exceeds c_4 or c_5), the more confident the eyewitness will be when identifying that face.

For notational purposes, let $\boldsymbol{x}$ be the set of memory signals generated by the faces in a given lineup. That is, $\boldsymbol{x} = \{x_1, x_2, x_3, \ldots x_k\}$, where the x_i are the memory signals generated by individual faces, with x_1 representing the memory signal generated by the suspect in the lineup, and k is lineup size. For the Independent Observations model, the decision variable used to decide whether to make a positive ID, $f(\boldsymbol{x})$, is the raw memory-match signal (x_i) of the face that generates the MAX signal. That is, for the Independent Observations model, $f(\boldsymbol{x}) = \max(\boldsymbol{x})$.

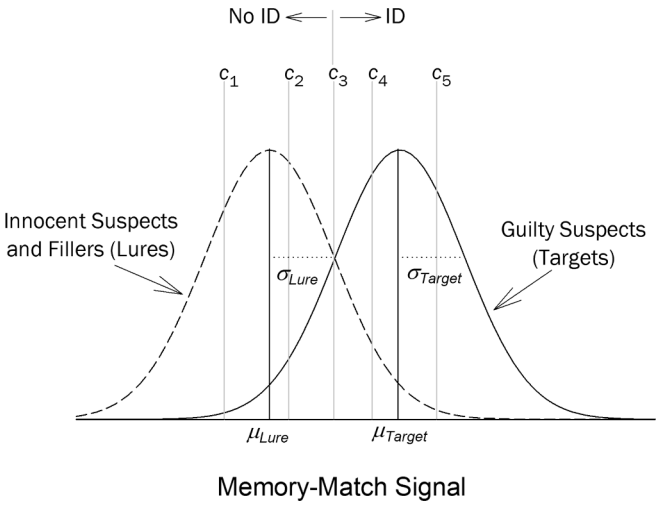

Fig. 2. This is the same model depicted in Fig. 1 except that the innocent suspect/filler distribution has been collapsed to a single distribution with a mean set to μ_{Lure}. In addition, confidence criteria have been superimposed on the raw (untransformed) memory-match signals because there are the memory signals that the Independent Observations model assumes are used to compare the MAX face to the confidence criteria (c_3 through c_5). The overall decision criterion is c_3.

The Ensemble model is much the same except that the raw memory signals depicted in Fig. 2 are all transformed by subtracting away the mean memory signal generated by the faces in the lineup. Conceptually, it is still a standard signal detection model like that depicted in Fig. 2, but the "memory match signal" is now conceptualized as the difference between the raw memory-match signal generated by a given face and the mean memory signal. This difference score will, on average, be greater for the guilty suspect in a target-present lineup than for fillers and innocent suspects.

According to this model, a strong memory-match signal (far to the right) exists not just when the raw signal for the MAX face is strong but when the difference between that raw signal and the mean memory signal is large. As with the Independent Observations model, only the MAX face is a candidate for being identified, but the decision variable is now $f(\boldsymbol{x}) = \max(\boldsymbol{x}) - \text{mean}(\boldsymbol{x})$. Note that $\text{mean}(\boldsymbol{x})$ represents the mean of all k faces in the lineup, including the MAX face. A reasonable alternative would be to subtract from $\max(\boldsymbol{x})$ the mean of the remaining $k-1$ faces in the lineup. This model turns out to be linearly related to the Ensemble model and is thus effectively the same model (Wixted et al., 2018).

If the $\max(\boldsymbol{x}) - \text{mean}(\boldsymbol{x})$ value exceeds c_3, a positive ID of the MAX face is made. Unlike the Independent Observations model, if $\max(\boldsymbol{x})$ is very strong in an absolute sense, a positive ID might not be made if the memory signals generated by all the faces in the lineup are also similarly strong.

Modeling lineup rejections

According to either model, if the decision variable falls *below* the overall criterion (c_3), the lineup is rejected, and that is the situation of interest here. When the lineup is rejected, confidence might still be based solely on the memory signal associated with the (unchosen) MAX face, with confidence being higher the weaker that signal happens to be. That is, even though the MAX face is not explicitly chosen, confidence in the lineup rejection might still be based on $f(\boldsymbol{x}) = \max(\boldsymbol{x})$ (Independent

Observations model) or $f(\boldsymbol{x}) = \max(\boldsymbol{x}) - \text{mean}(\boldsymbol{x})$ (Ensemble model), depending on which model is correct.

Fitting the MAX versions of each model to lineup rejections required some modification to the programs that have been used in the past to fit positive IDs, but the changes were straightforward. They were straightforward because no modifications to the previously reported likelihood functions for the Independent Observations and Ensemble models (Wixted et al., 2018) were needed to specify the MAX versions of these models for lineup rejections. The only issue that needed to be addressed is that—given maximum likelihood parameter estimates—the predicted confidence ratings for lineup rejections in which confidence is based on the guilty suspect's face (because it is the MAX face) or a filler's face (because it is the MAX signal) are not separately tracked. For example, a dataset might have 100 high-confidence positive IDs to a guilty suspect's face (i.e., the guilty suspect was the MAX face 100 times) and 25 high-confidence positive IDs to TP fillers (i.e., a TP filler was the MAX face 25 times), and it might also have 50 high-confidence lineup rejections. Unlike for high-confidence positive IDs, whether the MAX face was the guilty suspect or a TP filler is unknown for high-confidence lineup rejections. Because these two categories of lineup rejections cannot be disentangled in observed data, their corresponding predicted values (computed using the maximum likelihood parameter estimates) were aggregated together when fitting the models to the data.

Instead of relying on the MAX face when the lineup is rejected, confidence might be based on the average face-memory signal, with confidence being higher the weaker the AVG signal is. For a given lineup that has been rejected, the mean of the lineup memory signals is conceptualized as a random variable drawn from a distribution of means.

For the Independent Observations model, the mean decision variable for lineup rejections is computed when $f(\boldsymbol{x}) = \max(\boldsymbol{x})$ falls below c_3. Under those conditions, neither the mean nor the standard deviation of the distribution of means is independent of the lineup rejection decision outcome. As a result, the derivation of the relevant likelihood function is somewhat involved.

For the Ensemble model, the mean decision variable is computed when $f(\boldsymbol{x}) = \max(\boldsymbol{x}) - \text{mean}(\boldsymbol{x})$ falls below c_3, but this conditionality does not affect the mean and standard deviation of the relevant distribution of means. As a result, the derivation of the relevant likelihood function is much more straightforward. The Appendix provides the mathematical derivations of the likelihood functions corresponding to the AVG versions of the Independent Observations and Ensemble models. For those models, we assume that the decision variable switches from $f(\boldsymbol{x})$, which differs for the Independent Observations and Ensemble models, to $g(\boldsymbol{x}) = \text{mean}(\boldsymbol{x})$, regardless of which model is used to predict confidence in positive IDs.

Because both the Independent Observations and Ensemble models have both a MAX version and an AVG version for lineup rejections, there are four models in all. All four models include at least six parameters—μ_{Target} plus five confidence criteria—and three of the four models also include a parameter that captures the correlation between memory signals in the lineup (r). The mean and standard deviation of the lure distribution were defined to be 0 and 1, respectively, and an equal-variance model was assumed for simplicity. We fit all four models to five different lineup datasets, four from our lab and one from a different lab. The details of the fits are presented next, and the story turned out to be similar for each. Specifically, the fits of both models consistently support the idea that lineup rejections are based on the face that generates the MAX memory signal in the lineup, not on the AVG memory signal.

Method

The four models were fit to data from four different projects in our lab that focused on unrelated issues and sometimes included additional conditions that are not of interest here (e.g., a showup condition in which a single innocent or guilty suspect is presented). We refer to these datasets as Datasets A through D. As noted below, Datasets B and C have already been published, whereas Datasets A and D have not previously been reported. To test for generality, we also fit a dataset from a different lab (Brewer & Wells, 2006), and we refer to it as Dataset E. The Brewer and Wells paper is often cited in support of the claim that the confidence-accuracy relationship is weak for lineup rejections.

The experimental task was methodologically the same in all cases except that different stimulus materials were used, and lineup size varied between six and nine faces. In the standard lineup condition of each experiment, participants first watched a short mock-crime video involving a single perpetrator, completed a brief distractor task, and then made a recognition decision from a six-person simultaneous photo lineup (Datasets A, B, and D), a nine-person simultaneous photo lineup (Dataset C), or an eight-person simultaneous photo lineup (Dataset E).

In Datasets A through D, half the participants were randomly assigned to receive a target-present lineup, and the other half were randomly assigned to receive a target-absent lineup. In Dataset E, each participant watched two videos and was tested with a target-present lineup for one video and a target-absent lineup for the other. In all datasets, a target-present lineup consisted of a photo of the perpetrator from the mock-crime video plus five or more fillers, whereas a target-absent lineup consisted of six or more fillers.

For each participant, the fillers for Datasets A though D were randomly drawn from a large pool of possible filler photos (the same fillers were used for all lineups in Dataset E). The photos in the pool were selected to match the basic physical characteristics of the perpetrator (e.g., clean-shaven white male with short brown hair, approximately 20 years of age). Participants could select one face as being the perpetrator or reject the lineup by clicking the "Not Present" button. After their identification decision (e.g., identification or reject), the participant rated their confidence level (0 %-100 %). Each participant made only one or two recognition memory decisions (plus a confidence rating), so a relatively large number of participants was tested (via Amazon Turk for Datasets A through D and via undergraduate and community groups for Dataset E).

Results

Dataset A: These data were taken from the standard six-person simultaneous lineup condition of an experiment comparing that condition to two other conditions (a showup condition consisting of only one test face, and a rate-them-all condition in which a confidence rating was made to every face in a six-person lineup). For model-fitting purposes, the confidence ratings were collapsed into low (0–60), medium (70–89), and high (90–100) bins. This method of collapsing is common because doing so creates confidence bins with similar numbers of observations. In addition, having only three bins requires only three free parameters to estimate the confidence criteria, which helps control the overall number of free parameters that must be estimated for a given model fit. Table 1 presents the raw frequency counts for the various lineup decisions made with low, medium, or high confidence. The number of participants tested with a target-present lineup (N_{TP}) was 1271, and the number of participants tested with a target-absent lineup (N_{TA}) was 1334, bringing the total N to 2605. For target-present lineups, the hit rate (number of suspect IDs divided by the number of target-present lineups) was .74, the filler ID rate (number of filler IDs divided by the number of target-

Table 1
Frequency counts for Dataset A.

	Target Present			Target Absent	
Confidence	Suspect	Filler	Reject	Filler	Reject
Low	222	52	106	245	280
Med	314	28	73	97	323
High	409	16	51	56	333

Table 2
Maximum likelihood parameter estimates, number of free parameters (npar), and chi-square goodness-of-fit statistics for each model fit to Dataset A.

Model	μ_{Target}	c_1	c_2	c_3	c_4	c_5	r	npar	χ^2
Ind Obs MAX	2.15	0.57	1.03	1.42	1.98	2.57	0.53	7	31.03
Ind Obs AVG	2.24	−0.33	0.04	1.52	2.05	2.63	0.30	7	37.12
Ens MAX	2.42	0.90	1.18	1.44	1.87	2.37	–	6	52.51
Ens AVG	2.42	−0.36	0.51	1.44	1.87	2.38	0.66	7	52.43

present lineups) was .08, and the lineup rejection rate (number of lineup rejections divided by the number of target-present lineups) was .18. For target-absent lineups, the filler ID rate was .30, and the lineup rejection rate was .70.

The four models (i.e., the MAX and AVG versions of the Independent Observations model and the MAX and AVG versions of the Ensemble model) were fit to the data shown in Table 1 using maximum likelihood estimation. Table 2 shows the estimated parameter values and the chi-square goodness-of-fit statistics.

With regard to the Independent Observations model, both the MAX and AVG decision-variable versions had 7 free parameters, but the MAX version provided a somewhat better fit ($\chi^2 = 31.03$ vs. $\chi^2 = 37.12$). With regard to the Ensemble model, the MAX and AVG decision-variable versions provided nearly identical fits ($\chi^2 = 52.51$ vs. $\chi^2 = 52.43$). However, the AVG version had one additional free parameter (r) because the MAX version does not include a correlation parameter.[2] Moreover, setting r to 0 for the AVG version (reducing the number of free parameters for that version to 6) dramatically worsened the fit, $\chi^2(1) = 92.88 - 52.43 = 40.45$, $p < .001$. Thus, this parameter was essential, and given the close chi-square goodness-of-fit values for the two versions of the model, any penalty applied for the extra parameter in the AVG version would likely render the MAX version of the Ensemble model the winner. Indeed, both AIC and BIC for the MAX version (9006.74 and 9041.93, respectively) were lower than the corresponding value for the average version (9008.54 and 9049.60, respectively). Therefore, according to the Ensemble model as well, there is no reason to favor the average decision variable over the MAX decision variable for lineup rejections.

Although the purpose of this investigation was not to distinguish between the Independence Observations model vs. the Ensemble model, it is worth noting that the Independence Observations model provided a noticeably better fit to this dataset. However, as noted earlier, Wixted et al. (2018) previously argued that goodness-of-fit may not be the best way to distinguish between these two models. First, the Ensemble-MAX model has one fewer free parameter than Independent Observations MAX model. Second, when simulated data are generated using parameters similar to what is often observed in real data, the Independent Observations model has a much easier time fitting data generated by the Ensemble model than vice versa (Shen et al., 2023; Wixted et al., 2018). In other words, the Independent Observations model is the more flexible of the two. Thus, the best way to differentiate between them is to test their a priori theoretical predictions (see Shen et al., 2023). Still, for the present results, the goodness-of-fit advantage for the Independent Observations model is larger than it usually is, so it seems fair to say that, if anything, the results favor it over the Ensemble model.

[2] When the MAX rule is used for the Ensemble model, the subtractive process eliminates information about the correlation in much the same way that a within-subjects t-test is based on a dependent variable in which correlated error variance has been subtracted away.

Interestingly, for all of the models that included a correlation parameter (three of the four models), the fit was improved significantly by allowing its value to be positive. Of course, this is as it should be as faces in a lineup are, by design, included because they share a certain number of features (and are features that will match memory of the perpetrator). Even so, this is the first clear model-based evidence supporting the existence of correlated memory signals in lineups.

Dataset B: These data come from Experiment 1 of Yilmaz et al. (2022). That paper also reported an exact replication of Experiment 1, and we have combined the data from the original and exact replication experiments for model-fitting purposes. Table 3 presents the raw frequency counts for the various lineup decisions made with low (0–60), medium (70–89), or high confidence (90–100). For this experiment, $N_{TP} = 631$ and $N_{TA} = 567$, bringing the total N to 1198. For target-present lineups, the hit rate was .76, the filler ID rate was .06, and the lineup rejection rate was .18. For target-absent lineups, the filler ID rate was .30, and the lineup rejection rate was .70.

As before, the four models (two versions of the Independent Observations model and two versions of the Ensemble model) were fit to the data shown in Table 3 using maximum likelihood estimation. Table 4 shows the estimated parameter values and the chi-square goodness-of-fit statistics. With regard to the Independent Observations model, the AVG and MAX decision-variable versions provided nearly identical fits ($\chi^2 = 21.97$ vs. $\chi^2 = 21.29$, respectively), with a very slight edge going to the MAX version. With regard to the Ensemble model, the average and MAX decision-variable versions also provided nearly identical fits ($\chi^2 = 20.70$ vs. $\chi^2 = 21.64$, respectively), but the AVG version had an extra free parameter (r). Setting its value to 0 once again dramatically worsened the fit, $\chi^2(1) = 51.03 - 20.70 = 30.35$, $p < .001$, so the inclusion of this free parameter was essential. Once the difference in the number of free parameters is considered, the edge goes to the MAX version again. That is, both AIC and BIC for the MAX version (4108.50 and 4139.03, respectively) were lower than the corresponding values for the AVG version (4109.22 and 4144.84, respectively). Therefore, as with Dataset A, there is no compelling reason to favor the AVG decision variable over the MAX decision variable for lineup rejections, though it is a much closer call for this dataset.

Dataset C: These data come from Experiment 2 of Yilmaz et al. (2022), which involved a nine-person simultaneous photo lineup. Table 5 presents the raw frequency counts for the various lineup decisions made with low, medium, or high confidence. For this

Table 3
Frequency counts for Dataset B.

	Target Present			Target Absent	
Confidence	**Suspect**	**Filler**	**Reject**	**Filler**	**Reject**
Low	123	27	47	95	111
Med	153	5	39	55	146
High	203	5	29	18	142

Table 4
Maximum likelihood parameter estimates, number of free parameters (npar), and chi-square goodness-of-fit statistics for each model fit to Dataset B.

Model	μ_{Target}	c_1	c_2	c_3	c_4	c_5	r	npar	χ^2
Ind Obs MAX	2.13	0.48	0.99	1.37	1.95	2.57	0.62	7	21.29
Ind Obs AVG	2.25	−0.40	0.04	1.50	2.06	2.68	0.21	7	21.97
Ens MAX	2.48	0.91	1.21	1.46	1.91	2.44	–	6	21.64
Ens AVG	2.49	−0.46	0.73	1.46	1.91	2.45	0.58	7	20.70

Table 5
Frequency counts for Dataset C.

	Target Present			Target Absent	
Confidence	Suspect	Filler	Reject	Filler	Reject
Low	56	14	30	50	61
Med	62	7	17	16	66
High	66	1	6	10	40

experiment, $N_{TP} = 259$ and $N_{TA} = 243$, bringing the total N to 502. For target-present lineups, the hit rate was .71, the filler ID rate (number of filler IDs divided by the number of target-present lineups) was .08, and the lineup rejection rate (number of lineup rejections divided by the number of target-present lineups) was.20. For target-absent lineups, the filler ID rate was .31, and the lineup rejection rate was .69.

Table 6 shows the estimated parameter values and the chi-square goodness-of-fit statistics for the maximum-likelihood fits of the relevant models to the data presented in Table 5. With regard to the Independent Observations model, the MAX version provided a much better fit than the AVG version, ($\chi^2 = 8.43$ vs. $\chi^2 = 31.77$, respectively). With regard to the Ensemble model, the MAX and AVG versions provided similar fits ($\chi^2 = 12.28$ vs. $\chi^2 = 10.34$), with the edge going to the AVG version. Setting r to 0 equalized the number of free parameters for the two versions of the Ensemble model, but it again significantly worsened the fit, $\chi^2(1) = 15.02 - 10.34 = 4.68$, $p < .05$. Thus, as with the two previous datasets, this correlation parameter was necessary to provide a good fit. Moreover, once the difference in the number of free parameters is considered, the edge goes to the MAX version once again. That is, both AIC and BIC for the MAX version (1755.98 and 1781.29, respectively) were lower than the corresponding value for the AVG version (1757.12 and 1786.65, respectively).

Dataset D: The experiment from which these data were taken had two standard simultaneous lineup conditions, a short exposure condition and a long exposure condition, to which participants were randomly assigned. As might be expected, overall performance was better in the long-exposure condition, so we fit the models to the data from each condition separately. Consider first the data from the short-exposure condition.

Table 7 presents the raw frequency counts for the various lineup decisions made with low, medium, or high confidence. For this condition, $N_{TP} = 874$ and $N_{TA} = 879$, bringing the total N to 1753. For target-present lineups, the hit rate was .56, the filler ID rate was .22, and the lineup rejection rate was also .22. For target-absent lineups, the filler ID rate was .48, and the lineup rejection rate was .52.

Table 8 shows the estimated parameter values and the chi-square goodness-of-fit statistics for the maximum-likelihood fits of the models to the data shown in Table 7. With regard to the Independent Observations model, the MAX and AVG versions provided nearly identical fits ($\chi^2 = 19.17$ vs. $\chi^2 = 19.80$, respectively), and the same was true for the Ensemble model ($\chi^2 = 31.69$ vs. $\chi^2 = 30.35$). Fixing r at 0 for the AVG version of the Ensemble model to equalize the number of free parameters with the MAX version at 6 significantly worsened the fit, $\chi^2(1) = 44.73 - 30.35 = 14.38$, $p < .001$. Thus, as in the previous datasets, the AVG version needed r to fit the data, and once the difference in the number of free parameters is taken into account, the edge goes to the MAX version of the Ensemble model yet again. That is, both AIC and BIC for the MAX version (6583.82 and 6616.63, respectively) were lower than the corresponding value for the AVG version (6584.36 and 6622.64, respectively).

Next consider first the data from the long-exposure condition. Table 9 presents the raw frequency counts for the various lineup decisions made with low, medium, or high confidence. For this condition, $N_{TP} = 929$ and $N_{TA} = 887$, bringing the total N to 1816. For target-present lineups, the hit rate was .72, the filler ID rate was.11, and the lineup rejection rate was .17. For target-absent lineups, the filler ID rate was .39, and the lineup rejection rate was .61.

Table 10 shows the estimated parameter values and the chi-square goodness-of-fit statistics for the maximum-likelihood fits of the models to the data shown in Table 9. With regard to the Independent Observations model, the MAX version provided a better fit than the AVG version ($\chi^2 = 4.66$ vs. $\chi^2 = 9.76$, respectively). With regard to the Ensemble model, the AVG version outperformed the MAX version in terms of the unadjusted chi-square goodness-of-fit statistic ($\chi^2 = 16.43$ vs. $\chi^2 = 19.32$, respectively), though the AVG version needed the extra r parameter to win that competition. That is, eliminating r in the AVG

Table 7
Frequency counts for Dataset D (short exposure).

	Target Present			Target Absent	
Confidence	Suspect	Filler	Reject	Filler	Reject
Low	177	123	100	253	175
Med	145	44	49	109	162
High	169	27	40	63	117

Table 6
Maximum likelihood parameter estimates, number of free parameters (npar), and chi-square goodness-of-fit statistics for each model fit to Dataset C.

Model	μ_{Target}	c_1	c_2	c_3	c_4	c_5	r	npar	χ^2
Ind Obs MAX	2.44	1.00	1.34	1.66	2.17	2.72	0.31	7	8.43
Ind Obs AVG	2.16	−0.50	−0.18	1.52	2.17	2.78	0.11	7	31.77
Ens MAX	2.30	0.78	1.11	1.42	1.92	2.46	–	6	12.28
Ens AVG	2.29	−0.53	0.26	1.42	1.92	2.45	0.29	7	10.34

Table 8
Maximum likelihood parameter estimates, number of free parameters (npar), and chi-square goodness-of-fit statistics for each model fit to Dataset D (short exposure).

Model	μ_{Target}	c_1	c_2	c_3	c_4	c_5	r	npar	χ^2
Ind Obs MAX	1.48	0.32	0.73	1.10	1.79	2.30	0.47	7	19.17
Ind Obs AVG	1.54	−0.56	−0.19	1.19	1.84	2.35	0.12	7	19.80
Ens MAX	1.64	0.72	0.95	1.18	1.67	2.09	–	6	31.69
Ens AVG	1.64	−0.60	0.26	1.18	1.67	2.09	0.44	7	30.35

Table 9
Frequency counts for Dataset D (long exposure).

	Target Present			Target Absent	
Confidence	Suspect	Filler	Reject	Filler	Reject
Low	110	39	71	169	130
Med	187	37	47	119	188
High	372	23	43	55	226

version by fixing it value at 0 significantly worsened the fit, $\chi^2(1) = 33.48 - 16.43 = 17.05$, $p < .001$. This time, penalizing the AVG version for its extra parameter yielded a split decision. With regard to AIC, the AVG version still provided the better fit (6390.08 vs. 6391.00 for the average and MAX versions, respectively). With regard to BIC, the MAX version provided the better fit (6428.61 vs. 6424.03 for the AVG and MAX versions, respectively).

Dataset E: These data were taken from an experiment reported by Brewer and Wells (2006). Not only are these data from an independent lab, but they are often cited in support of the claim that the confidence-accuracy relationship is weak for lineup rejections. Thus, if the asymmetry in confidence-accuracy relationships for positive and negative IDs from lineups is the result of different decision variables being used, these findings may offer the best chance of detecting that fact.

In this study, subjects first watched a video in which they viewed two targets, a thief and a waiter. For each condition, $N_{TP} = 600$ and $N_{TA} = 600$, bringing the total N to 1200. All subjects were tested for their ability to identify the thief from an 8-member simultaneous lineup. After completing the lineup memory test for the thief, the subjects were subsequently tested for their ability to identify the waiter from a different 8-member simultaneous lineup. Thus, because each subject was tested twice, there were 2400 observations in all. Table 11 presents the raw frequency counts. Collapsed across the Thief and Waiter conditions, for target-present lineups, the hit rate was .49, the filler ID rate was .20, and the lineup rejection rate was .31. For target-absent lineups, the filler ID rate was .44, and the lineup rejection rate was .56.

The confidence-accuracy relationships for positive and negative IDs (averaged over the thief and waiter conditions) are shown in Fig. 3. Note that, for positive IDs, the confidence-accuracy relationship in Fig. 3A is plotted in the conventional way, with accuracy (% Correct) quantifying the accuracy of suspect IDs (i.e., filler IDs are excluded from the calculation). The data for positive IDs are typically plotted this way because it answers the relevant legal question: Given that a suspect was identified with a particular level of accuracy, how likely is that ID to be accurate (Wixted & Wells, 2017)? The relationship is stronger for positive IDs (and high-confidence accuracy is much higher for positive IDs than for negative IDs), but a relationship for negative IDs is nevertheless apparent.

Smith et al. (2023) hypothesized that a focus on suspect IDs for positive IDs may explain the asymmetry in the confidence-accuracy relationship for positive vs. negative IDs. Unlike for positive IDs, for negative IDs, an outcome is counted as correct or incorrect whether the MAX signal is generated by the suspect or a filler because, when a lineup is rejected, it is not known which face generated the MAX signal. To make the plots for positive and negative IDs more comparable, in Fig. 3B, accuracy for positive IDs was re-computed by counting any ID from a TP lineup as correct (a suspect ID or a filler ID), whereas any ID from a TA lineup was counted as being incorrect (a suspect ID or a filler ID). As illustrated in Fig. 3B, it remains the case that the confidence-accuracy relationship is stronger for positive IDs, and a high-

Table 11
Frequency counts for Dataset E (Brewer & Wells, 2006).

		Target Present			Target Absent	
Condition	Confidence	Suspect	Filler	Reject	Filler	Reject
Thief	Low	30	35	71	53	47
	Med	50	36	85	73	110
	High	142	34	118	71	245
Waiter	Low	56	46	24	107	71
	Med	96	44	28	131	76
	High	215	42	48	91	125

Table 10
Maximum likelihood parameter estimates, number of free parameters (npar), and chi-square goodness-of-fit statistics for each model fit to Dataset D (long exposure).

Model	μ_{Target}	c_1	c_2	c_3	c_4	c_5	r	npar	χ^2
Ind Obs MAX	2.07	0.65	1.04	1.32	1.75	2.30	0.45	7	4.66
Ind Obs AVG	2.08	−0.38	−0.16	1.33	1.76	2.31	0.16	7	9.76
Ens MAX	2.28	0.90	1.13	1.32	1.64	2.10	–	6	19.32
Ens AVG	2.27	−0.15	0.53	1.32	1.64	2.09	0.30	7	16.43

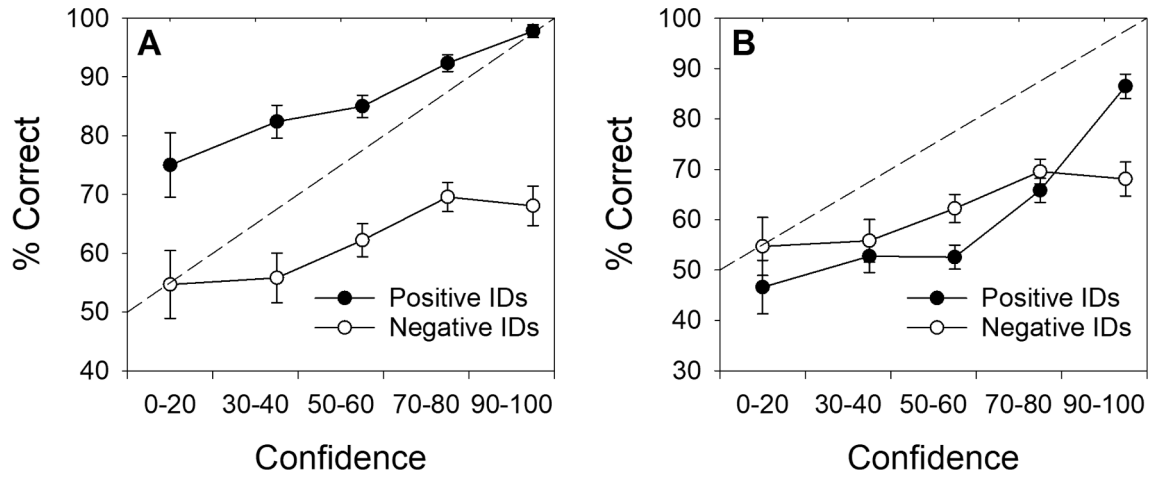

Fig. 3. Confidence-accuracy data from Brewer and Wells (2006) after averaging across the Thief and Waiter conditions. The data were also collapsed over two between-subjects experimental conditions (namely high-vs.-low-similarity foils, and biased vs. unbiased instructions). A. The accuracy score for positive IDs is based on suspect IDs only. B. The accuracy score for positive IDs is based on suspect or filler IDs (with filler IDs counted as correct for TP lineups and incorrect for TA lineups). The dashed line in each plot does not represent perfect calibration (where 0% confidence represents 0% accuracy and 100% confidence represents 100% accuracy) but instead represents a perfect confidence-accuracy relationship (where 0% confidence represents chance accuracy of 50% correct and 100% confidence represents perfect performance, or 100% accuracy).

confidence positive ID is much more accurate than a high-confidence negative ID. The question of interest here is whether that difference arises because confidence in a lineup rejection is based on an AVG signal.

Table 12 shows the estimated parameter values and the chi-square goodness-of-fit statistics for the maximum-likelihood fits of the MAX and AVG versions of the Ensemble and Independent Observations models to the data. The data from the thief and waiter conditions were fit separately and then the results were averaged together. With regard to the Independent Observations model, the MAX version provided a much better fit than the AVG version ($\chi^2 = 12.97$ vs. $\chi^2 = 26.34$, respectively), but the Ensemble model returned the opposite verdict before correcting for the differing number of free parameters ($\chi^2 = 14.45$ vs. $\chi^2 = 8.97$). Once again, penalizing the average version for its extra parameter yielded a split decision. With regard to AIC, the AVG version provided an ever-so-slightly better fit (6487.22 vs. 6487.41 for the average and MAX versions, respectively). With regard to BIC, the MAX version provided the better fit (6522.85 vs. 6517.95 for the AVG and MAX versions, respectively).

Thus, on balance, the verdict would have to favor the MAX decision variable. Stated differently, it would hard to make a compelling case in favor of the AVG decision variable based on these findings.

General discussion

The idea that the decision variable for lineup rejections might be based on an average memory signal was first suggested by Weber and Brewer (2006):

Alternatively, as a negative decision indicates that the stimulus does not match well with any of the relevant items in memory, confidence in negative decisions could be based on the average (or median) match between all the relevant items in memory and the test stimulus. This type of aggregated basis for confidence therefore suggests a potential difference between confidence in positive and negative decisions that could underlie the observed positive–negative calibration difference (p. 19).

Lindsay et al. (2013) considered this possibility as well, as did Yilmaz et al. (2022). This hypothesis seems plausible because when a lineup is rejected, no single face is identified; instead, the entire set of faces is collectively rejected. Simulations conducted by Yilmaz et al. (2022) suggested that part of the explanation for the asymmetric confidence-accuracy relationship might be that the decision variable used to rate confidence a lineup is rejected is the average of the memory signals generated by the faces in the lineup. However, Yilmaz et al. (2022) did not attempt to directly test that hypothesis, as we have done here.

The model-fitting approach we used required modifying the

Table 12
Maximum likelihood parameter estimates, number of free parameters (npar), and chi-square goodness-of-fit statistics for each model fit to Dataset E (averaged over the Thief and Waiter conditions).

Model	μ_{Target}	c_1	c_2	c_3	c_4	c_5	r	npar	χ^2
Ind Obs MAX	1.57	0.91	1.19	1.37	1.66	2.08	0.18	7	12.97
Ind Obs AVG	1.60	−0.24	0.09	1.45	1.72	2.12	0.01	7	26.34
Ens MAX	1.71	1.12	1.30	1.43	1.64	1.97	–	6	14.45
Ens AVG	1.70	0.05	0.87	1.43	1.64	1.97	0.28	7	8.97

likelihood functions for the Independent Observations and Ensemble models to allow for the possibility that confidence for lineup rejections is based on an average memory signal. However, when those newly derived models were fit to empirical data from multiple simultaneous lineup experiments, a relatively clear verdict was obtained. For the Independent Observations model, the MAX version fit better than the average version in the clear majority of comparisons. The verdict was similar for the Ensemble model. However, depending on how the difference in the extra parameter associated with AVG model was addressed (AIC or BIC), the AVG version of the model sometimes yielded a better fit. Still, our overall findings favor the idea that the MAX memory signal determines confidence not only for positive IDs but also for negative IDs. Also, as noted earlier, it seems fair to say that the idea that the MAX signal determines confidence in a lineup rejection is the default view—the idea that an average signal might be used as the basis of confidence was proposed only in response to an empirical anomaly (namely, the comparatively weak confidence-accuracy relationship for lineup rejections). Thus, even if the AVG model had slightly outperformed the MAX model across the totality of these datasets, we would not have considered that outcome to be sufficient evidence to overturn the default view. Since AVG model did not even perform that well, there is even less reason to adopt a new perspective.

At the same time, our model-fitting results do not prove that the AVG model is wrong. Going forward, more direct tests might help to establish its viability. For example, a standard simultaneous lineup condition could be compared to a condition in which witnesses who reject the lineup are asked to provide a confidence rating to everyone in the lineup. In the standard lineup condition, when the witness rejects the lineup, the question would be "How certain are you that the person from the video is not in this lineup?" This rating would apply to the collective set of faces in the lineup. For the rate-them-all condition, the faces would be individually rated, and for each one, the question would be "How certain are you that this is not the person from the video?" For each participant in the rate-them-all condition, we would have both an average rating and a MAX rating. The question of interest is whether the distribution of collective ratings from the standard condition (based either on the MAX or AVG signal) more closely resembles the distribution of MAX ratings or the distribution of average ratings from the rate-them-all condition. Still, until more direct evidence in its favor is adduced, the assumption that an AVG decision variable underlies confidence in lineup rejections should not replace the default view.

One interesting issue that emerged for the first time is that, when fitting signal detection models to lineup data, the results consistently indicated that the competing memory signals in lineups are correlated. This means that if one face in the lineup generates a weak memory signal, all of the faces in the lineup tend to do the same. This is expected given that a lineup contains faces that were selected precisely because they are similar to each other, so the memory signals they generate should ebb and flow together (Wixted et al., 2018; Shen et al., 2023). Still, in past research involving fits of the Independent Observations model, the estimated correlation parameter did not differ from 0 (e.g., Shen et al., 2023).[3] In Shen et al. (2023), this result likely occurred because the similarity of fillers was manipulated across conditions, and discriminability increased monotonically as filler similarity decreased. The Independent Observations model most clearly predicts this filler-similarity pattern when the correlation parameter equals 0, with the magnitude of the filler-similarity effect decreasing as the correlation increases. Hence, the best fit was obtained when the correlation parameter was 0 even though the correlation must increase with increasing filler similarity. One reason why Shen et al. (2023) argued in favor of the Ensemble model was that it more naturally accounts for the filler-similarity findings.

In the datasets analyzed here, we fit models to data from individual conditions, and the expected correlation was finally reliably detected by both the Independent Observations model and the AVG version of the Ensemble model. However, our results leave unexplained the mystery that the averaging hypothesis was originally advanced to explain: why is the confidence-accuracy relationship for positive vs. negative IDs often asymmetrical? An attractive but ultimately untenable explanation would appeal to a similar asymmetry observed in the list-learning literature, where the variance of the target distribution is found to be greater than the lure distribution almost invariably. Mickes et al. (2011) argued that this asymmetry may explain why the confidence-accuracy relationship is typically weaker for "new" decisions compared to "old" decisions—even in the list-learning paradigm. However, a similar asymmetry is typically not observed when memory is tested using lineups, and it sometimes goes in the opposite direction (e.g., Shen et al., 2023). Thus, a different explanation for the asymmetry sometimes observed for lineups presumably applies.

An approach that may unravel the mystery would be to investigate the underlying mechanisms that give rise to the memory signals that signal detection theory takes for granted. The signal detection models under consideration make assumptions about those memory signals (e.g., they are normally distributed, the effective signal might be the MAX signal minus the mean signal, etc.), but they are silent about the mechanisms that give rise to them in the first place. Recently, Colloff et al. (2021) and Shen et al. (2023) proposed a simplified feature-matching mechanism that generates the face recognition memory signal, and much more comprehensive feature-matching models have been used to guide thinking about recognition for some time (e.g., Shiffrin & Steyvers, 1997). Yet, so far, those models do not offer reasons as to why the confidence-accuracy relationship for lineup rejections would differ from the confidence-accuracy relationship for positive IDs.

Other feature-matching models might offer some insight, such as the global similarity model advanced by Mewhort and Johns (2000). Global similarity based on feature matching is still assumed to contribute to the memory signal, but Mewhort and Johns (2000) found that the rejection of novel items was enhanced when test items contained novel features. This was true even when the remaining features strongly matched a studied item, yielding a strong familiarity signal based on overall similarity. They called this the "extralist feature effect" (see Osth et al., 2023), and it is akin to what others call "recall to reject" (e.g., Rotello & Heit, 2000). Yet, even this approach does not seem to account for the weak confidence-accuracy relationship for lineup rejections. To the extent that the extralist feature effect occurs (e.g., if all of the faces in the lineup have a feature *not* shared by the representation of the perpetrator in the brain), one might expect the lineup rejection to be made both confidently and accurately. But the empirical puzzle to be explained is the differentially low accuracy associated with high-confidence lineup rejections.

Although lineup rejections remain a bit of a mystery, it seems that confidence in those decisions is based on the MAX face, just as positive IDs are. Thus, the take-home message of our investigation is that when a lineup is rejected, the weaker the decision variable associated with the MAX face is, the more confident the witness is that the perpetrator is not in the lineup.

CRediT authorship contribution statement

Anne S. Yilmaz: Conceptualization, Methodology, Writing – original draft, Writing – review & editing. **John T. Wixted:** Conceptualization, Formal analysis, Writing – review & editing, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial

[3] The standard version of the Ensemble model does not have a correlation parameter, so fits of this model would not detect the correlation even if it is present. The AVG version of the model used here for the first time also detected correlated memory signals.

interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Appendix

Investigating the possibility that lineup rejections are based not on the MAX signal but instead on a different decision variable, $g(\boldsymbol{x}) = \text{mean}(\boldsymbol{x})$, requires modifying the likelihood functions that have been used to fit the models in the past. We next specify the likelihood functions for this alternative model of confidence in lineup rejections, first in general terms and then in model-specific terms (i.e., in terms specific to the Ensemble model and then in terms specific to the Independent Observations model).

Lineup rejections based on the average memory signal (in general terms)

The likelihood functions for both models consist of the joint probability of multiple events. For example, in the case of a lineup rejection on a given trial, there is (1) the probability of observing a given memory strength, x_i, for face i, (2) the probability that x_i is the MAX value in the lineup, (3) the probability that the decision variable for positive IDs, $f(\mathbf{x})$, falls below the decision criterion given that x_i is the MAX value, and (4) the probability that the decision variable for negative IDs, $g(\mathbf{x})$, falls above a confidence criterion given that $f(\mathbf{x})$ falls below the decision criterion.

More formally, assuming a standard signal detection model, the probability of observing target memory strength x_i (event 1) is given by a Gaussian distribution with mean, μ_i and standard deviation σ:

$$P(x_i) = \phi(z_i) \tag{1}$$

where ϕ is the Gaussian probability density function and $z_i = \frac{x_i - \mu_i}{\sigma}$. As a concrete example, assume that it is a target-present lineup and that the face in question is that of the guilty suspect such that $x_i = x_1$ and $\mu_1 = \mu_{Target}$. In that case, $P(x_1)$ is the probability of drawing a particular memory strength signal from the target distribution in Fig. 2.

Continuing with this example (i.e., $x_i = x_1$), consider next the probability that x_1 is the MAX signal in the lineup. The probability that x_1 is greater than the value of *all* fillers in a lineup of size k (event 2) is:

$$P(x_2 \cdots x_k < x_1) = \prod_{j=2}^{k} \Phi\left(\frac{x_1 - \mu_j}{\sigma}\right) \tag{2}$$

where Φ is the Gaussian CDF (i.e., the standard cumulative normal distribution). $x_2 \cdots x_k$ in this example correspond to the $k-1$ fillers in the lineup, so μ_j can be set to 0 for convenience. The quantity $\Phi\left(\frac{x_1 - \mu_j}{\sigma}\right)$ represents the probability of drawing a value less than x_1 for filler j, and the product from $j = 2$ to k in Equation (2) is the probability that all $k-1$ fillers fall below x_1, in which case $x_1 = \max(\boldsymbol{x})$.

In our running example, $x_i = x_1$ (this is the suspect's memory signal) and $x_1 = \max(\boldsymbol{x})$. In addition, $f(\mathbf{x})$ represents the decision variable for positive IDs, which always involves the MAX signal but differs for the two models. That is, $f(\mathbf{x})$, is equal to x_1 according to the Independent Observations model and is instead equal to $x_1 - \text{mean}(\boldsymbol{x})$ according to the Ensemble model. The probability that the decision variable associated with x_1, $f(\mathbf{x})$, falls below the decision criterion (event 3) is simply:

$$P(f(\boldsymbol{x}) < c_3 | x_1 = \max(\boldsymbol{x})) \tag{3}$$

where c_3 is the overall decision criterion in Fig. 2.

For lineup rejections, the decision variable is $g(\boldsymbol{x}) = \text{mean}(\boldsymbol{x})$. The probability that $g(\boldsymbol{x})$ falls *above* a relevant confidence criterion (c_i) for lineup rejections (event 4) given that $x_1 = \max(\boldsymbol{x})$ and that $f(\mathbf{x})$ falls below c_3 is given by:

$$P(g(\boldsymbol{x}) > c_i | x_1 = \max(\boldsymbol{x}), f(x) < c_3) \tag{4}$$

where $g(\boldsymbol{x}) = \text{mean}(\boldsymbol{x})$, and c_i is c_1 or c_2 in Fig. 2.

Thus, the probability of observing x_1 (i.e., the target in a target-present lineup in our running example) *and* the probability that x_1 is greater than the value of all fillers (i.e., lures) in a lineup of size k *and* the probability that the decision variable for making a positive ID, $f(\boldsymbol{x})$, falls below the decision criterion (c_3), *and* the probability that the decision variable for rating confidence in a negative ID, $g(\boldsymbol{x})$, falls above c_i is given by Equation (1) × Equation (2) × Equation (3) × Equation (4).

Lineup rejections based on $g(\boldsymbol{x})$ according to the Ensemble model

The details for Equations (1), 2, and 3 have been presented before (Wixted et al., 2018), but the details of Equation (4) are new and are presented here for the first time. For the Ensemble model, the model-specific version of Equation (4) is simple and straightforward, so we begin there. For a given lineup that has been rejected, the mean of $\boldsymbol{x}$ is conceptualized as a random variable drawn from a distribution of means. Thus, we need to specify the mean and standard deviation of that distribution. For a single lineup with k faces, $\text{mean}(\boldsymbol{x}) = (1/k)\sum_1^k x_i$. For a target-present lineup, the memory signal for the guilty suspect is drawn from a normal distribution with a mean of μ_{Target} and a standard deviation of σ, whereas the memory signals for the fillers are drawn from a normal distribution with a mean of μ_{Lure} and a standard deviation of σ. That is, $x_{i=1}\ N(\mu_{Target}, \sigma)$ and $x_{i\neq 1}\ N(\mu_{Lure}, \sigma)$. Thus, the mean of means across target-present lineups of size k is equal to $\frac{\mu_{Target} + (k-1)\mu_{Lure}}{k}$. For target-absent lineups, the mean of means is equal to $\frac{k\mu_{Lure}}{k} = \mu_I$. Because we set $\mu_{Lure} = 0$ for convenience, the mean of means for target-present and target-absent lineups come to $\frac{\mu_{Target}}{k}$ and 0, respectively. For the uncorrelated case ($r = 0$), the standard deviation for the mean of means is, in both cases, equal to $\sigma/\sqrt{k}$, where σ is set to 1 for convenience. Thus, according to the central limit theorem, for target-present lineups, $\overline{X}_i\ N\left(\frac{\mu_{Target}}{k}, \frac{1}{\sqrt{k}}\right)$, and for target-absent lineups, $\overline{X}_i\ N\left(0, \frac{1}{\sqrt{k}}\right)$. However, as noted earlier, the memory signals generated by the faces in a lineup are likely correlated ($r > 0$), and in that case the standard deviation for the mean of

means is given by $\frac{\sqrt{1+r(k-1)}}{\sqrt{k}}$.

It is worth highlighting the fact that, ordinarily, the correlation coefficient does not show up in the equations for the Ensemble model even when the memory signals in a lineup are assumed to be correlated. The reason is that when the decision variable is $\max(\boldsymbol{x}) - \text{mean}(\boldsymbol{x})$, as it is for MAX version of the Ensemble model for both positive and negative IDs (and as it still is for positive IDs even for the average version of lineup rejections under consideration now), correlated error is subtracted out and therefore cannot be estimated from the data (i.e., the correlation coefficient is not usually a free parameter for this model). However, if the decision variable switches to $\text{mean}(\boldsymbol{x})$ when a lineup is rejected (the average version), now the correlation can be estimated as a free parameter even for the Ensemble model because, for lineup rejections, the correlation has not been subtracted out of the decision variable. Thus, this version of the Ensemble model has one additional free parameter (r) compared to the MAX version that assumes a $\max(\boldsymbol{x}) - \text{mean}(\boldsymbol{x})$ decision variable for both positive and negative IDs.

In more detail, for lineup i that has been rejected, if $\overline{X}_i$ falls below c_1, the lineup is rejected with high confidence. If it falls above c_1 but below c_2, the lineup is rejected with medium confidence, and if it falls above c_2, the lineup is rejected with low confidence. What makes these equations so straightforward and easy to use in the case of the Ensemble model is that even though the mean decision variable is relevant only when $f(\boldsymbol{x}) = \max(\boldsymbol{x}) - \text{mean}(\boldsymbol{x})$ falls below c_3 (i.e., only when the lineup is rejected), that conditionality does not affect the mean and standard of the relevant distribution of means. This is true because both the mean and standard deviation of the distribution of means are independent of the variable that determines the decision outcome, namely, $\max(\boldsymbol{x}) - \text{mean}(\boldsymbol{x})$. Thus, for the Ensemble model, event 4 is

$$P(\overline{X}_i) = \phi(Z_i)$$

Where $Z_i = \frac{\overline{X}_i - \mu_M}{\sigma_M}$, with μ_M representing the mean of means ($\frac{\mu_G}{k}$ for target-present lineups and 0 for target-absent lineups) and σ_M representing the standard deviation of means ($\frac{1}{\sqrt{k}}$ for both lineup types in the uncorrelated case and $\frac{\sqrt{1+r(k-1)}}{\sqrt{k}}$ in the more likely correlated scenario).

Lineup rejections based on $g(\boldsymbol{x})$ according to the Independent Observations model

The situation is more complicated for the Independent Observations model, where the mean decision variable for lineup rejections is computed when $f(\boldsymbol{x}) = \max(\boldsymbol{x})$ falls below c_3. Under those conditions, the mean and standard deviation of the distribution of means are not independent of the decision outcome. Instead, when the lineup is rejected, the k memory signals in the lineup from which the mean is computed are conceptualized as having been drawn from a truncated normal distribution ranging from a minimum of $-\infty$ to a maximum of $\max(\boldsymbol{x})$. Under such conditions, the distribution of means would not be Gaussian, and the mean and standard deviation of that distribution could not be directly computed based on the central limit theorem, as was the case for the Ensemble model. This raises a question: When specifying this mean (i.e., the hypothesized decision variable) as a random variable for a given rejected lineup with a given $\max(\boldsymbol{x})$, what distribution is the mean value drawn from? This is the complication associated with modeling confidence in a lineup rejection based on an average memory signal according to the Independent Observations model.

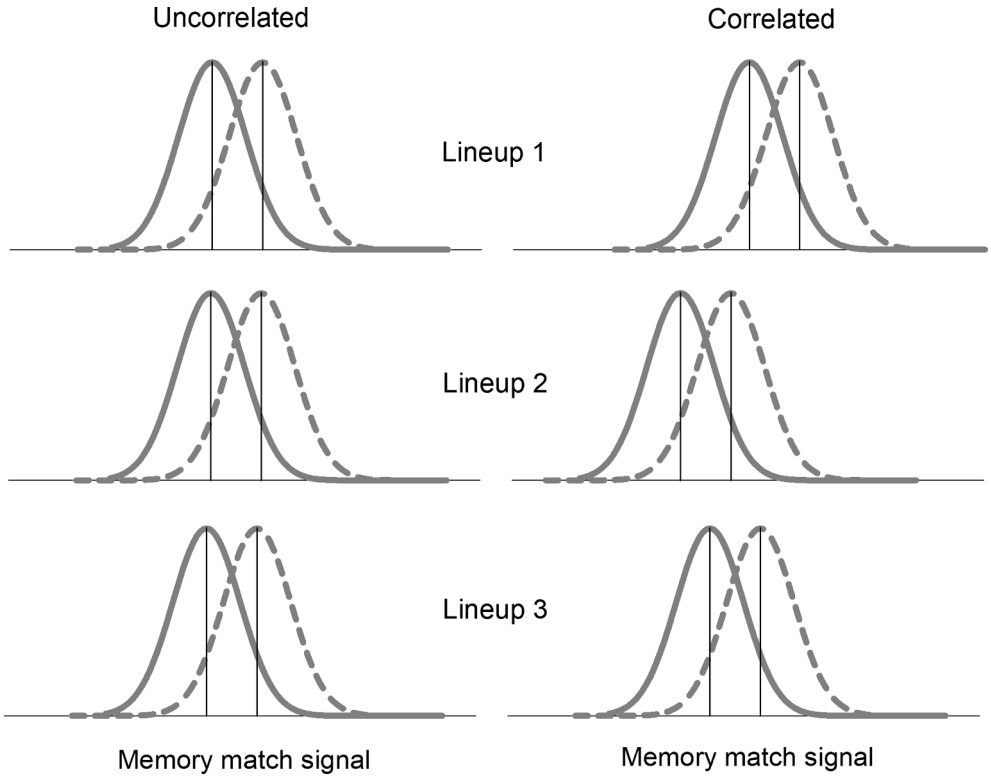

Fig. A1. An illustration uncorrelated (left column) and correlated (right column) memory signals across three lineups. In the left column, between lineup variance (σ_b^2) is equal to zero. In the right column, σ_b^2 is greater than zero. The larger σ_b^2 is relative to within lineup variance (σ_w^2), the more the memory signals are correlated. The magnitude of the correlation (r) is equal to $r = \frac{\sigma_b^2}{\sigma_b^2+\sigma_w^2}$.

Fortunately, a nearly exact approximation is available. For a given value of $\max(\boldsymbol{x})$, equations to compute the mean and variance of single values randomly drawn from the corresponding truncated normal distribution-that is, with values drawn below $\max(\boldsymbol{x})$-have been provided (see Greene, 2003, p. 759). From there, it is a simple matter to compute the mean and standard deviation of the mean of k values randomly drawn from truncated target or lure normal distributions. For a given $\max(\boldsymbol{x})$, we denote the mean and standard deviation of the distribution of means as μ_T and σ_T, respectively, where the subscript T indicates that the parameter is based on values drawn from a truncated normal distribution. For a target-present lineup, these values are based on $k-1$ draws from the lure distribution truncated at $\max(\boldsymbol{x})$ and one draw from the target distribution that is also truncated at $\max(\boldsymbol{x})$. For a target-absent lineup, these values are based on k draws from the lure distribution truncated at $\max(\boldsymbol{x})$.

With μ_T and σ_T in hand, even though the distribution of means is not Gaussian in form, we can use the Gaussian probability density function as a close approximation to estimate the probability of drawing a particular mean, $\overline{X}_i$, given that the lineup was rejected:

$$P(\overline{X}_i) = \phi(Z_i)$$

where $Z_i = \frac{\overline{X}_i - \mu_T}{\sigma_T}$. The Gaussian PDF approximation becomes more precise the larger k is according to the central limit theorem. But even with $k = 6$ (a standard lineup size and one that we used in most of the research reported here), the approximation is surprisingly close to being exact. For the Independent Observations model, this is event 4 specified by Equation (4) above.

Finally, we need to incorporate correlated memory signals into the Independent Observations model. Although this was simple and straightforward for the Ensemble model (requiring only a modification to the equation for the standard deviation of the distribution of means), more than that is required for the Independent Observations model, as illustrated in Fig. A1. The left panel illustrates three lineups in which memory signals are uncorrelated, whereas the right panel illustrates three lineups in which the memory signals are positively correlated. When memory signals are uncorrelated, the variance in the memory signals generated by guilty suspects and fillers reflect random error within lineups ($\sigma^2 = \sigma_w^2$), with no additional variance occurring between lineups. By contrast, when memory signals are correlated, it means that when the memory signal generated by the guilty suspect is strong, the memory signals generated by the fillers are also strong, and when the memory signal generated by the guilty suspect is weak, the memory signals generated by the fillers are also weak. In other words, the variability in memory signals has a between-lineup component (σ_b^2). This represents an additional source of variability between lineups such that $\sigma^2 = \sigma_w^2 + \sigma_b^2$. The larger σ_b^2 is relative to σ_w^2, the more correlated the memory signals are, with the correlation (r) being equal to $r = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_w^2}$. Because we set $\sigma^2 = 1$ throughout, this means that $\sigma_w^2 + \sigma_b^2 = 1$, so the equation for r simplifies to $r = \sigma_b^2$.

For modeling purposes, a positive correlation is introduced to the likelihood function for the Independent Observations model by adding another event, which, in this case, is another random variable to create between-lineup variance. To do so, δ is drawn from a Gaussian distribution with a mean of 0 and standard deviation of σ_b, and it is added to the means of both the target and lure distributions (thereby creating the kind of variability observed in the right column of Fig. A1). More formally, $\delta\ N(0, \sigma_b)$, and this can be conceptualized as event 0 (occurring prior to events 1 through 4). Thus, the probability of observing target memory strength x_i (event 1) is now given by a Gaussian distribution with mean, μ_i and standard deviation σ:

$$P(x_i) = \phi(z_i)$$

where, now, $z_i = \frac{x_i - (\mu_i + \delta_i)}{\sigma}$. As before, for the guilty suspect in a target-present lineup, $\mu_i = \mu_{Target}$ (an estimated parameter) and for all other lineup members in target-present or target-absent lineups, $\mu_i = \mu_{Lure} \equiv 0$.

In the case of correlated memory signals for the Independent Observations model, across all five events (events 0 through 4), there are three random variables, with each integrated from $-\infty$ to $+\infty$: δ_i, x_i, and $\overline{X}_i$. The triple integral makes for a slow fitting of this version of the model, but the fit is nonetheless precise.

Summary. Both versions of the Independent Observations model (i.e., versions that assume a MAX or average decision variable for lineup rejections) have the same seven free parameters: μ_{Target}, c_1, c_2, c_3, c_4, c_5, and r. However, the two corresponding versions of the Ensemble model do not both have seven free parameters. The version of the Ensemble model that assumes a $\max(\boldsymbol{x}) - \text{mean}(\boldsymbol{x})$ decision variable for both positive and negative IDs has six free parameters (all but r), but the version of the Ensemble model that assumes a $\max(\boldsymbol{x}) - \text{mean}(\boldsymbol{x})$ decision variable for positive IDs and average decision variable for negative IDs has seven free parameters (now including r). All four versions of the models under consideration here (two versions of the Independent Observations and two versions of the Ensemble model) were verified using model recovery simulations. That is, the models differentially fit their own simulated data very accurately, and the maximum likelihood fits precisely estimate the programmed parameter values.

References

Akan, M., Robinson, M. M., Mickes, L., Wixted, J. T., & Benjamin, A. S. (2021). The effect of lineup size on eyewitness identification. *Journal of Experimental Psychology: Applied, 27*(2), 369–392.

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.

Colloff, M. F., Wilson, B. M., Seale-Carlisle, T. M., & Wixted, J. T. (2021). Optimizing the selection of fillers in police lineups. *Proceedings of the National Academy of Sciences, 118*, e2017292118; 10.1073/pnas.2017292118.

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. 10.1002/acp.3178Hanczakowski, M., Butowska, E., Beaman, C. P., Jones, D. M., & Zawadzka, K. (2021). The dissociations of confidence from accuracy in forced-choice recognition judgments. *Journal of Memory and Language, 117*, 104189.

Jou, J., Flores, S., Cortes, H. M., & Leka, B. G. (2016). The effects of weak versus strong relational judgments on response bias in Two-Alternative-Forced-Choice recognition: Is the test criterion-free? *Acta Psychologica, 167*, 30–44.

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.

Mewhort, D. J. K., & Johns, E. E. (2000). The extralist-feature effect: Evidence against item matching in short-term recognition memory. *Journal of Experimental psychology: General, 129*(2), 262–284. https://doi.org/10.1037/0096-3445.129.2.262

Mickes, L., Hwe, V., Wais, P. E., & Wixted, J. T. (2011). Strong memories are hard to scale. *Journal of Experimental Psychology: General, 140*, 239–257.

Miyoshi, K., Kuwahara, A., & Kawaguchi, J. (2018). Comparing the confidence calculation rules for forced-choice recognition memory: A winner-takes-all rule wins. *Journal of Memory and Language, 102*, 142–154.

Osth, A. F., Zhou, A., Lilburn, S. D., & Little, D. R. (2023). Novelty rejection in episodic memory.*Psychological. Review, 130*(3), 720–769. https://doi.org/10.1037/rev0000407

Rotello, C. M., & Heit, E. (2000). Associative recognition: A case of recall-to-reject processing. *Memory & Cognition, 28*(6), 907–922. https://doi.org/10.3758/BF03209339

Shen, K. J., Colloff, M. F., Vul, E., Wilson, B. M., & Wixted, J. T. (2023). Modeling face similarity in police lineups. *Psychological Review, 130*(2), 432–461. https://doi.org/10.1037/rev0000408

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

Weber, N., & Brewer, N. (2006). Positive Versus Negative Face Recognition Decisions: Confidence, Accuracy, and Response Latency. *Applied Cognitive Psychology, 20*(1), 17–31.

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., Rydell, S. M., & Seelau, E. P. (1993). The selection of distractors for eyewitness lineups. *Journal of Applied Psychology, 78*(5), 835–844. 10.1037/0021-9010.78.5.835Wells, 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. 10.1023/A:1025750605807.

Wixted, J. T., Vul, E., Mickes, L., & Wilson, B. W. (2018). Models of lineup memory. *Cognitive Psychology, 105*, 81–114.

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.

Zawadzka, K., Higham, P. A., & Hanczakowski, M. (2017). Confidence in forced-choice recognition: What underlies the ratings? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 43*(4), 552–564.

ACKNOWLEDGEMENTS

Chapter 2, in full, is a reprint of the material as it appears in: Yilmaz, A.S**.** & Wixted, J.T. (2024). What latent variable underlies confidence in lineup rejections? *Journal of Memory and Language*, *135,* 104493. The dissertation author was the primary researcher and author of this paper. A full reprint of this material within a dissertation is allowed without permission by Elsevier, the publisher of *Journal of Memory and Language*, given that the dissertation author is also the author of the original paper, and the dissertation is not published commercially. Professor John Wixted, who is both a co-author and the dissertation committee chair, has given permission to use this work in fulfillment of my dissertation requirements.

Received: 2 December 2023 | Revised: 15 March 2024 | Accepted: 5 April 2024

DOI: 10.1002/acp.4196

RESEARCH ARTICLE

WILEY

Response bias modulates the confidence-accuracy relationship for both positive identifications and lineup rejections in a simultaneous lineup task

Anne S. Yilmaz | Xiaoqing Wang | John T. Wixted

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

Correspondence
John T. Wixted, Department of Psychology, University of California, San Diego, La Jolla, CA 92093, USA.
Email: jwixted@ucsd.edu

Funding information
UCSD Yankelovich Center; Center of Academic Research and Training in Anthropogeny (CARTA)

Abstract

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

KEYWORDS
confidence-accuracy relationship, lineup rejections, signal detection theory

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

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

However, the strength of the confidence-accuracy relationship appears to be much less impressive when it comes to lineup rejections. Indeed, in contrast to the strong relationship for positive IDs, the relationship between confidence and accuracy for lineup rejections is often (but not always) found to be negligible (Arndorfer & Charman, 2022; Brewer & Wells, 2006). Thus, a high-confidence lineup rejection is not necessarily indicative of high accuracy like it is in the case of a positive identification.

Although the field has already reached a de facto consensus about the nature of the confidence-accuracy relationship in the case of lineup rejections, no formal review of the past literature has been conducted in the manner previously done for positive IDs by Wixted and Wells (2017). We therefore did so here by reviewing the confidence-accuracy relationship for lineup rejections reported in 12 experiments

(Brewer et al., 2002; Brewer & Wells, 2006; Carlson et al., 2017; Dobolyi & Dodson, 2013; Dodson & Dobolyi, 2016; Horry et al., 2012; Keast et al., 2007; Palmer et al., 2013; Sauer et al., 2008, 2010; Sauerland & Sporer, 2009; Weber & Brewer, 2004). Except for a few studies that did not report confidence for lineup rejections, these are the same experiments reviewed by Wixted and Wells (2017) to assess the confidence-accuracy relationship for positive IDs made using a 100-point confidence scale. The data sets span a variety of study designs, such as: simultaneous and sequential lineups, same-race and cross-race identifications, methodologies (e.g., disconfirmation and reflection, immediate presentation and delayed presentation, etc.), as well as different sample populations (e.g., adults and children). Figure 1 shows the average confidence-accuracy relationship for positive suspect IDs reported by Wixted and Wells (2017) and for lineup rejections. Clearly, the relationship is weaker for lineup rejections.

The question of interest here is why there is an asymmetry between the confidence-accuracy relationships for positive versus lineup rejections. Picking up on an idea suggested by Brewer and Wells (2006) and Lindsay et al. (2013), Yilmaz et al. (2022) hypothesized that confidence in lineup rejections might be determined by the average memory signal because no singular face is identified when a lineup is rejected. This is in contrast to positive IDs, where confidence is presumably based on the one that generates the strongest memory-match signal (the MAX face). However, using a model-fitting approach across six different data sets, Yilmaz and Wixted (2024) found that confidence in a lineup rejection also appears to be based on the MAX face (i.e., the less familiar the MAX is, the more confidence the witness is in rejecting the lineup).

Smith et al. (2023) hypothesized that a focus on suspect IDs for positive IDs may explain the asymmetry in the confidence-accuracy relationship for positive versus lineup rejections. Unlike for positive IDs, for lineup rejections, an outcome is counted as correct or incorrect whether the MAX signal is generated by the suspect or a filler because, when a lineup is rejected, it is not known which face generated the MAX signal. Moreover, the distribution of memory-match signals for innocent and guilty suspects overlap to a lesser degree (i.e., discriminability is higher) compared to the distribution of MAX memory-match signals (regardless of whether the MAX face is the suspect or a filler). However, for positive IDs, the confidence-accuracy relationship is not appreciably affected over a fairly wide range of discriminability, so it is not clear that this factor would explain the asymmetry.

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

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

In eyewitness research, the focus is often on finding ways to induce more conservative responding for witnesses as it reduces the likelihood of a misidentification (Clark, 2005). Common examples of

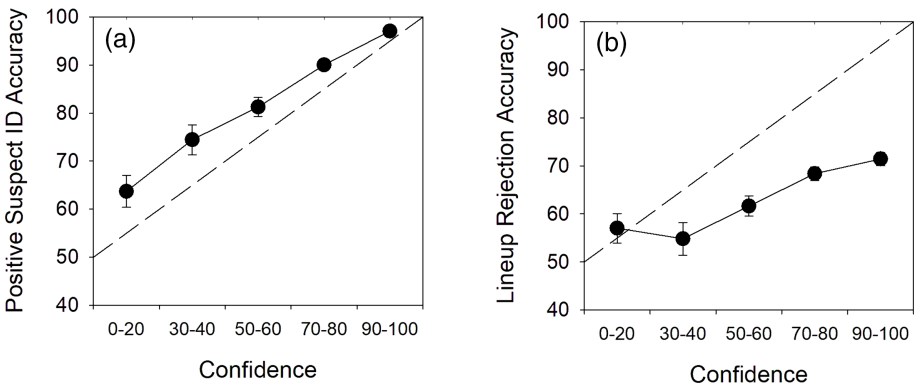

FIGURE 1 (a) Confidence-accuracy characteristic for positive IDs reported by Wixted and Wells (2017). (b) Confidence-accuracy characteristic for lineup rejections from the same studies that reported confidence for lineup rejections.

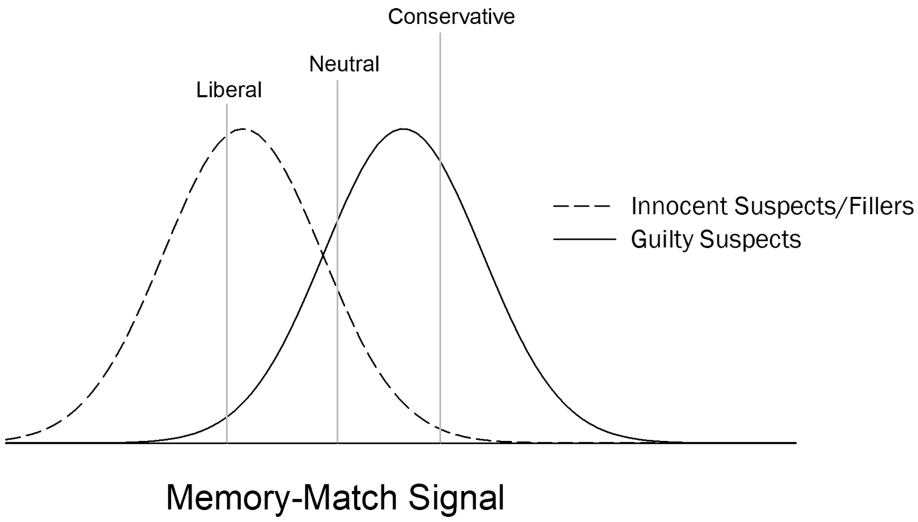

FIGURE 2 A standard signal detection model illustrating different place of the overall decision criterion for making positive IDs (liberal, neutral, and conservative).

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

1 | EXPERIMENT 1

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

2 | METHOD

2.1 | Participants

We recruited participants from Amazon's MTurk ($n = 2250$). All participants passed attention check questions and reported that they had not seen the stimulus video before. Participants were compensated 25 or 50 cents for their time. The participants included 42.8% Male (1006), 52% Female (1222), 0.25% Other (6), 0.08% Decline to Answer (2), and 4.85% no response (114). The ethnicity distribution of the participants was: 82.5% Caucasian (1939), 3.14% African-American (74), 8.42% Asian (198), 3.19% Latino (75), 0.5% Native-American (12), 0.26% Middle-Eastern (6), 0.12% Pacific-Islander (3), 1% Other (23), 0.6% Decline to Answer (13), and 0.3% No Response (7).

2.2 | Design and materials

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

2.3 | Procedure

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

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

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

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

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

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

3 | RESULTS

The overall choosing rate from TP lineups in the liberal condition (suspect IDs plus filler IDs divided the number of TP lineups) was .88, whereas the corresponding value for the conservative condition was .74, a difference that was significant, $\chi^2 = 33.94, p < .001$. The overall choosing rate from TA lineups in the liberal condition (filler IDs divided the number of TA lineups) was .42, whereas the corresponding value for the conservative condition was .26, a difference that was also significant, $\chi^2 = 33.65, p < .001$. In other words, choosing rates were significantly lower in the conservative condition for both TA and TP lineups, indicating that response bias was successfully manipulated.

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

Figure 3 presents the receiver operating characteristic (ROC) data. An ROC is a plot of the hit rate (suspect IDs from TP lineups divided by the number of TP lineups) versus the false alarm rate (estimated suspect IDs from TA lineups divided by the number of TA lineups) for three different decision criteria. Because there was no designated innocent suspect, the false ID rates were estimated by dividing the TA filler ID rates by lineup size (6). The left most point for each condition only counts suspect IDs made with high confidence, the middle point counts suspect IDs made with medium or high confidence, and the rightmost point counts suspect IDs made with low, medium, or high confidence. The rightmost points represent what is ordinarily considered to be the overall hit and false alarm rates, and it is visually apparent that it falls farther to the right in the liberal condition (reflecting the more liberal response bias). The two curves trace out essentially the same trajectory, indicting similar levels of discriminability (i.e., the response bias manipulation did not have the unintended consequence of differentially affecting discriminability). This is consistent with an earlier study by Mickes et al. (2017), which found that although discriminability was lower for the liberal and conservative conditions relative to an unbiased condition, they were similar to each other.

TABLE 1 Frequency counts by confidence bin for Experiment 1.

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

Abbreviations: TP(S) = suspect IDs from target-present lineups, TP(F) = filler IDs from target-present lineups, and TA = filler IDs (Positive) and lineup rejections (Reject) from target-absent lineups.

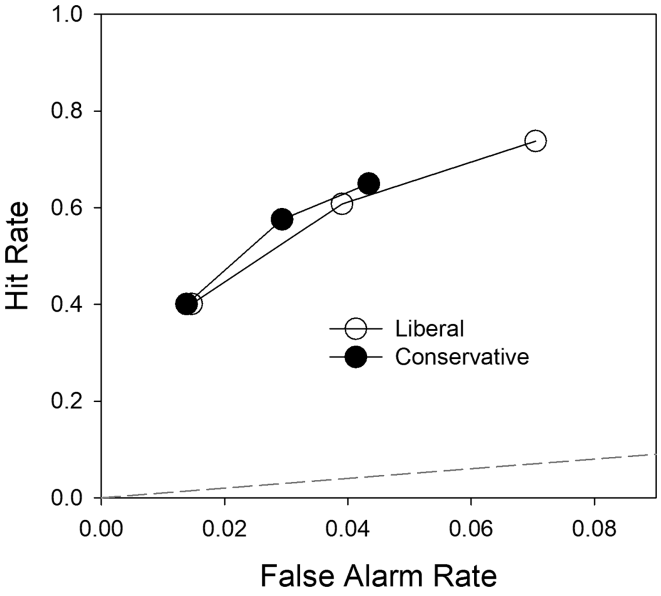

FIGURE 3 ROC data from the liberal and conservative conditions of Experiment 1. The dashed line represents chance performance.

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

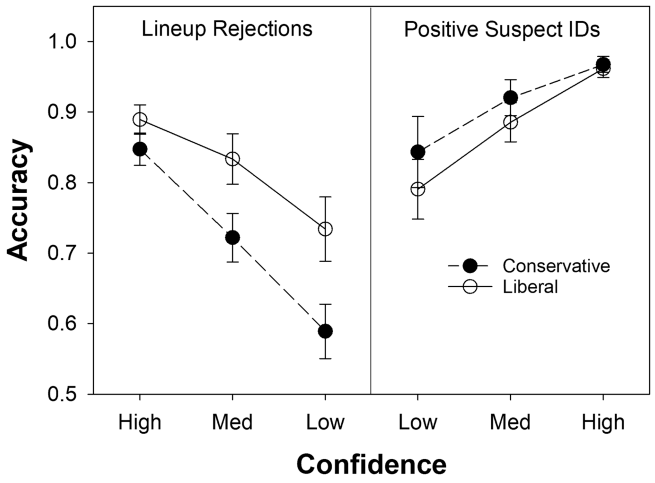

FIGURE 4 (Left panel) Confidence-accuracy characteristic for lineup rejections in the conservative and liberal conditions. (Right panel) Confidence-accuracy characteristic for positive IDs in the conservative and liberal conditions. The scale on the x-axis can be conceptualized as a 6-point confidence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

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

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

Within each confidence level considered individually (low, medium, high), pairwise comparisons for positive suspect IDs did not differ significantly for the conservative and liberal conditions. For lineup rejections, accuracy within confidence levels was significantly lower in the conservative condition (relative to the liberal condition) for low and medium confidence, $\chi^2(1) = 5.4, p = .014$, and $\chi^2(1) = 4.56, p = .033$, respectively. By contrast, the difference for high-confidence lineup rejections was not significant, $\chi^2(1) = 1.81, p = .178$.

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

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

4 | EXPERIMENT 2

In Experiment 2, everything was the same as in Experiment 1 except that we switched to a forced-choice procedure. Now, participants were always asked to choose the one lineup member who was most likely to be the perpetrator from the crime video. In addition, for the identified individual, participants were also asked to rate their confidence on a−100 to +100 scale, where −100 indicated complete certainty that the identified individual was not the perpetrator, and + 100 indicated complete certainty that the identified individual was the perpetrator (0 represented complete uncertainty).

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

4.1 | Participants

We recruited participants from Amazon's MTurk ($n = 3023$). We excluded 106 people due to having seen the stimulus video before. This left 2917 participants in the final analysis. Participants were compensated 25 or 50 cents for their time. The participants included 41.68% Male (1214), 57.08% Female (1665), 0.65% Other (19), 0.51% Decline to Answer (15), and 0.14% no response (4). The ethnicity distribution of the participants was: 76.69% Caucasian (2237), 8.98% African-American (262), 5.93% Asian (173), 5.07% Latino (148), 0.48% Native-American (14), 0.21% Middle-Eastern (6), 0.14% Pacific-Islander (4), 1.44% Other (42), 0.51% Decline to Answer (15), and 0.55% no response (16).

4.2 | Design and materials

The study was a randomized 2 (standard simultaneous vs. 6AFC simultaneous) × 2 (target present vs. target absent) design. The experiment

used the same mock-crime stimulus video and 45-second distractor task as above.

4.3 | Procedure

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

After clicking the "Next" button, participants then received one of two lineup conditions, either for a standard simultaneous lineup or a 6AFC simultaneous lineup. The standard simultaneous lineup had two rows of three photographs. In the target present condition, one photo in the lineup was of the guilty suspect while the other five photographs were fillers. The target absent condition did not contain a photo of the perpetrator. Instead, the lineup included a sixth filler photo. Filler photos in the lineup were randomly selected from a pool of 60 possible fillers, all description-matched to the guilty suspect. Participants could select a photograph as being the man from the video or they could reject the lineup by indicating that the man from the video was not present. At the top of the lineup, an instruction read, "Below is a lineup that *may or may not* contain the perpetrator from the video. If you believe that the perpetrator is present, please select his face. Otherwise, please click "Not Present" below." After participants selected a face or rejected the lineup, they give their confidence (1–100; 1 = completely unsure; 100 = completely sure).

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

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

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

5 | RESULTS

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

Figure 5 presents the ROC data for the two conditions of Experiment 2. For the Standard condition, the points represent positive suspect IDs. As is typical of lineup ROC data, it is not possible to plot suspect ID (hit) rates for lineup rejections because no face is identified when a lineup is rejected. For the 6AFC condition, by contrast, participants identified the MAX face and supplied a confidence rating even when the lineup was rejected. ROC points for lineup rejections could therefore be plotted even for rejections. That is, for TP lineups, it was known when the MAX rejected face was the guilty suspect (making it possible to plot the "hit rate" even when the face was technically rejected) and for TA lineups, the innocent suspect would be the identified MAX face 1/6 of the time. The ROC points for positive IDs and lineup rejections for the 6AFC condition are connected by a dotted

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

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

Abbreviations: TP(S) = suspect IDs from target-present lineups, TP(F) = filler IDs from target-present lineups, and TA = filler IDs (Positive) and lineup rejections (Reject) from target-absent lineups.

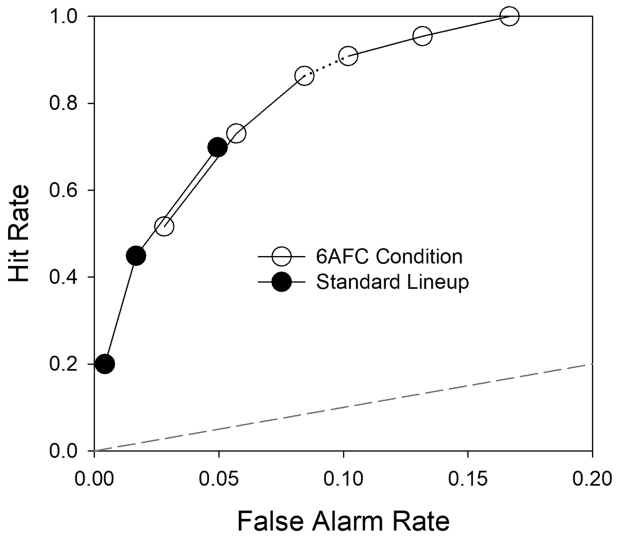

FIGURE 5 ROC data from the Standard Lineup and 6AFC Condition of Experiment 2. For the 6AFC condition, the leftmost three ROC points (open circles) represent positive IDs (as do the filled circles for the standard condition), whereas the rightmost three ROC points connected by a dotted line represent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

line to create one continuous ROC curve. As in Experiment 1, the two curves trace out essentially the same trajectory, indicting similar levels of discriminability (i.e., the 6AFC requirement did not have the unintended consequence of affecting discriminability relative to the standard condition). Instead, for positive IDs (the leftmost 3 points), the 6AFC condition resulted in a more liberal response bias. The effect was not problematic because our focus was on the slope of the CAC curves as response bias varied over a wide range.

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

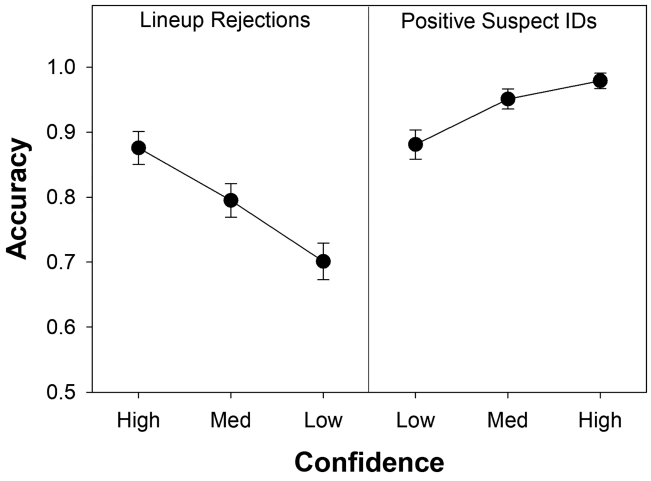

FIGURE 6 (Left panel). Confidence-accuracy characteristic for lineup rejections in the standard lineup condition of Experiment 2. (Right panel). Confidence-accuracy characteristic for positive IDs in the standard lineup condition of Experiment 2.

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

As shown in the right panel of Figure 7, the slopes for positive IDs for each decision criterion condition were ordered as predicted. That is, the slope of the confidence-accuracy relationship was steepest in the most liberal condition (−80) and shallowest in the most conservative condition (+80). Indeed, across all response bias conditions, the slopes were monotonically ordered (they became shallower as the response bias became more conservative).

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

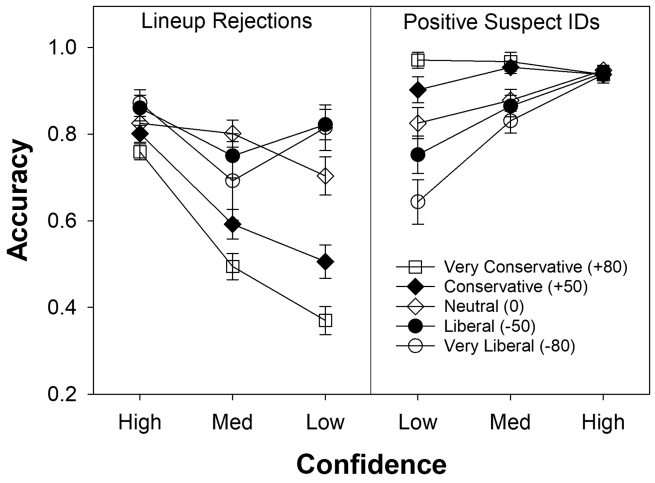

FIGURE 7 (Left panel). Confidence-accuracy characteristic for lineup rejections in the 6AFC condition of Experiment 2. (Right panel). Confidence-accuracy characteristic for positive IDs in the 6AFC condition of Experiment 2. The scale on the x-axis can be conceptualized as a 6-point confidence scale,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

values into the low- and high-confidence bins such that the frequency counts for each confidence level remained roughly equated. We then computed the two-point slopes for each response bias condition. For four out of the five response bias conditions, these two-point slopes for lineup rejections were ordered as predicted (whereas all five of the two-point slopes for positive IDs were ordered as predicted).

To determine how often this pattern of results for positive and lineup rejections would arise by chance, we computed a statistic consisting of the sum of squared differences between the predicted and observed rankings of slopes. For example, if the predicted order across the five conditions was 1, 2, 3, 4, 5, and if the observed order was 1, 2, 3, 5, 4 (the last two reversed relative to predictions, as in the lineup rejection data here), the statistic would be $(1-1)^2+(2-2)^2+(3-3)^2+(4-5)^2+(5-4)^2=2$. Next, we ran 10,000 bootstrap trials in which the observed rank order was randomly determined. For example, if the random order on a given bootstrap trial was 3, 1, 4, 5, 2, the bootstrap statistic for this trial would be $(1-3)^2+(2-1)^2+(3-4)^2+(4-5)^2+(5-2)^2=13$. wWe asked how often these randomly ordered bootstrap trials yielded a sum of squares statistic as small or smaller than the observed sum of squares statistics for positive IDs and lineup rejections separately. The result was significant for both positive IDs ($p=.008$) and lineup rejections ($p=.040$).

6 | GENERAL DISCUSSION

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

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

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

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

Two other factors, not investigated here, might also affect the strength of the confidence-accuracy relationship for lineup rejections. One factor is whether the decision variable itself might be causing the asymmetric empirical pattern of data shown earlier in Figure 2. Functionally for lineups, confidence for positive IDs is given in relation to a single face (i.e., the selected face, with the MAX memory signal). However, it is less clear what confidence is tied to for lineup rejections since the task for simultaneous lineups involves collectively rejecting a set of faces. Conceivably, confidence in lineup rejections is based on the average memory signal rather than on the MAX memory signal (as posited by Brewer & Wells, 2006; Lindsay et al., 2013; Yilmaz et al., 2022). The use of an average signal might yield a weaker confidence-accuracy relationship. However, recent research from our lab suggests this explanation may not be right. Using a model-fitting approach, we found evidence supporting the idea that confidence is

based on the MAX face regardless of whether a positive ID or a lineup rejection is made (Yilmaz & Wixted, 2024).

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

Whatever the explanation turns out to be, achieving a better understanding of the relationship between confidence and accuracy for lineup rejections seems important given that many of the DNA exoneration cases involving high-confidence misidentifications at trial began with something other than that (sometimes with a lineup rejection) on the initial test (Garrett, 2011). It is essential to focus on the results of the first test (Wells et al., 2020; Wixted et al., 2021), especially when the witness rejects the lineup, but a key question that has not yet been fully answered is when confidence informs accuracy for lineup rejections. The main finding reported here is that confidence in a lineup rejection is more informative when response bias is conservative compared to when it is liberal. Thus, if these results are confirmed by other labs using different stimulus materials, then for jurisdictions that use lineup instructions to encourage a conservative response bias, it would be safe to conclude that confidence in a lineup rejection has more information value than would otherwise be the case.

ACKNOWLEDGMENT

Supported by a grant from the UCSD Yankelovich Center and in part by the Center of Academic Research and Training in Anthropogeny (CARTA) Fellowship.

CONFLICT OF INTEREST STATEMENT

The authors have no conflicts of interest to declare.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in OSF at https://osf.io/w8hnd/?view_only=bc0463105dac4b76819d8d63399a026c.

ORCID

Anne S. Yilmaz https://orcid.org/0000-0003-4759-4910
John T. Wixted https://orcid.org/0000-0001-6282-5479

ENDNOTE

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

REFERENCES

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

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

Brewer, N., & Wells, G. L. (2006). The confidence-accuracy relationship in eyewitness identification: Effects of lineup instructions, foil similarity, and target-absent base rates. *Journal of Experimental Psychology: Applied*, *12*(1), 11–30. https://doi.org/10.1037/1076-898X.12.1.11

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

Clark, S. E. (2005). A Re-examination of the effects of biased lineup instructions in eyewitness identification. *Law and Human Behavior*, *29*(4), 395–424. https://doi.org/10.1007/s10979-005-5690-7

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

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

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

Horry, R., Palmer, M. A., & Brewer, N. (2012). Backloading in the sequential lineup prevents within-lineup criterion shifts that undermine eyewitness identification performance. *Journal of Experimental Psychology: Applied*, *18*(4), 346–360. https://doi.org/10.1037/a0029779

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

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

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

Mickes, L., Seale-Carlisle, T. M., Wetmore, S. A., Gronlund, S. D., Clark, S. E., Carlson, C. A., Goodsell, C. A., Weatherford, D., & Wixted, J. T. (2017). ROCs in eyewitness identification: Instructions versus confidence ratings. *Applied Cognitive Psychology*, *31*(5), 467–477. https://doi.org/10.1002/acp.3344

Palmer, M. A., Brewer, N., Weber, N., & Nagesh, A. (2013). The confidence-accuracy relationship for eyewitness identification decisions: Effects of exposure duration, retention interval, and divided attention. *Journal of Experimental Psychology: Applied*, *19*(1), 55–71. https://doi.org/10.1037/a0031602

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

Sauer, J., Brewer, N., Zweck, T., & Weber, N. (2010). The effect of retention interval on the confidence–accuracy relationship for eyewitness

identification. *Law and Human Behavior*, *34*(4), 337–347. https://doi.org/10.1007/s10979-009-9192-x

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

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

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

Weber, N., & Brewer, N. (2004). Confidence-accuracy calibration in absolute and relative face recognition judgments. *Journal of Experimental Psychology: Applied*, *10*(3), 156–172. https://doi.org/10.1037/1076-898x.10.3.156

Wells, G. L., Kovera, M. B., Douglass, A. B., Brewer, N., Meissner, C. A., & Wixted, J. T. (2020). Policy and procedure recommendations for the collection and preservation of eyewitness identification evidence. *Law and Human Behavior*, *44*(1), 3–36. https://doi.org/10.1037/lhb0000359

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

Wixted, J. T., 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*, *113*, 304–309.

Wixted, J. T., & Wells, G. L. (2017). The relationship between eyewitness confidence and identification accuracy: A new synthesis. *Psychological Science in the Public Interest*, *18*, 10–65.

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

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

Yilmaz, A. S., & Wixted, J. T. (2024). What latent variable underlies confidence in lineup rejections? *Journal of Memory and Language*, *135*, 104493.

How to cite this article: Yilmaz, A. S., Wang, X., & Wixted, J. T. (2024). Response bias modulates the confidence-accuracy relationship for both positive identifications and lineup rejections in a simultaneous lineup task. *Applied Cognitive Psychology*, *38*(2), e4196. https://doi.org/10.1002/acp.4196

ACKNOWLEDGEMENTS

Chapter 3, in full, is a reprint of the material as it appears in: Yilmaz, A.S., Wang, X., & Wixted, J.T. (2024). Response bias modulates the confidence-accuracy relationship for both positive IDs and lineup rejections in a simultaneous lineup task. *Applied Cognitive Psychology*, 38(2), e4196. The dissertation author was the primary researcher and author of this material. Permission to use the material as it appears was granted by John Wiley and Sons, the publisher of *Applied Cognitive Psychology*. All co-authors (Xiaoqing Wang and Professor John Wixted) and the dissertation committee chair (Professor John Wixted) have given permission to use this work in fulfillment of my dissertation requirements.

CONCLUSION

Although these studies focus on eyewitness identification (an applied domain), the three experiments proposed here seek to explain the asymmetry of confidence between positive IDs and rejections within the frameworks provided by basic memory science. Up until this point, it has been unclear why confidence is not particularly diagnostic of accuracy when a face is not recognized despite being highly diagnostic of accuracy when a face is recognized. Understanding the driving causes behind this asymmetry could prove to be invaluable. The two most recent consensus papers in the field outlined the importance of testing a witness' memory of a particular suspect only once, and they emphasized that a lineup rejection provides evidence of *innocence* instead of simply a lack of evidence of guilt (Wells et al., 2020; Wixted et al., 2021). Importantly, no subsequent test of memory yields more reliable information from the first. With that in mind, when we focus on understanding the first test of memory in exoneration cases involving eyewitness misidentifications at trial, we find a high frequency of initial lineup rejections made correctly by witnesses. For the majority of exoneration cases in which there is information available about the initial identification procedure, the witness did not confidently misidentify the innocent suspect on that test (and often correctly rejected them) despite being highly confident while making a misidentification at trial (Garrett, 2011; Yilmaz et al., 2024a).

These considerations underscore the importance of better understanding the information value of initial lineup rejections, but a puzzle has been why the confidence-accuracy relationship for lineup rejections is weak despite the relation being strong for positive identifications. One hypothesis as to why this asymmetry exists has to do with the

decision variable used in rejections. No support was found for participants using an AVG rule as the basis for confidence though (Yilmaz & Wixted, 2024). Even so, we did find that changing the task (i.e., the Reveal procedure) such that confidence was tied to a specific face instead of a set of faces for rejections affected the confidence-accuracy relationship (Yilmaz et al., 2022). Specifically, when we asked for a participant's belief that the suspect was *not* the perpetrator after they rejected the lineup (but before they gave their confidence rating), we were able to strengthen the confidence-accuracy relationship for rejections and also improve accuracy for high-confidence rejections compared to the standard simultaneous lineup.

As for why the confidence-accuracy relationship for lineup rejections is weak, it seems like a strong contributing factor may be response bias (Yilmaz et al., 2024b). The last chapter of this dissertation demonstrates how incrementally shifting the response criterion in any direction corresponds to a change in strength of relationship for *both* positive IDs and lineup rejections. At least with the standard simultaneous lineup, the ability to detect a relationship between confidence and accuracy for positive IDs comes at the cost of being less-able to detect a relationship between the two for lineup rejections.

Future directions of this research should aim to build upon procedures that enhance evidence of suspect innocence while maintaining the quality and quantity of information gathered about suspect guilt. Because the police can reasonably test a witness' memory of a given suspect only once using a lineup, increasing the amount of information gathered about a suspect on the first memory test is a worthy goal. One option would be to implement a rate-them-all procedure whenever the suspect is not identified, as proposed by Chapter 1 of this dissertation (Yilmaz et al., 2022) and was further investigated by Smith et

al. (2023). The benefit of this rate-them-all procedure is that the police can gather potentially exonerating confidence ratings on every face in the lineup (including the suspect), and it can also be administered in a double-blind fashion (Brewer et al., 2019). A drawback is that this type of design would be a significant departure from current police procedure, and therefore less likely to be adopted despite any benefits that the procedure brings. Also, recent research reported that a rate-them-all simultaneous lineup (administered on its own instead of after a standard simultaneous lineup) may introduce noise in decision-making process, which can lower discriminability (Yilmaz et al., 2024c). The impact of noise on decision-making is unclear in the design by Smith et al. (2023), and it may be the case that finding another variation that does not introduce as much noise would be more beneficial.

Another approach worth investigating is as follows: After a rejection decision is made by an eyewitness, police ask witnesses to select all of the faces in the lineup that they are sure are *definitely not* the perpetrator. This allows police to determine which faces are being rejected with high degrees of confidence, and which faces are not. One could analyze how confidence in this design relates to quantity and quality of evidence of innocence. This design would also allow the identity of the suspect to be hidden from the witness. It would also allow information to be gathered about the suspect (whether they were selected as a high-confidence rejection, or if they instead were unsure about the rejection of that particular face) while still maintaining the higher level of discriminability of the standard simultaneous lineup procedure. An advantage of this approach is that it would minimally depart from current police procedures and thus might be more feasible to implement in the real world.

Combining the two proposed experiments could also prove to be fruitful. Witnesses could select all of the faces in the lineup that they believe *definitely* did not commit the crime, but also indicate their confidence level for each of those selected faces. The additional confidence ratings could provide more information about the suspect to police (which could be a benefit), and may introduce less noise than if participants had to make a rating to every single face in the lineup.

It might also be worth exploring whether discriminability (*d'*) has any effect on the confidence-accuracy relationship for lineup rejections. In unpublished preliminary data, we were able to successfully manipulate discriminability by varying study time, but its effect on the confidence-accuracy relationship remained inconclusive. We plan on exploring this relationship more directly by implementing stimulus videos that use no-, low-, medium-, and high-blur to the perpetrator's face. It could be that rejections made with lower discriminability yield less diagnostic information than rejections made with higher degrees of discriminability because this pattern tends to show up in signal-detection-based simulations.

Regardless of the specific approach future research takes when trying to better understand lineup rejections, the need for this research remains clear. There have been three recent exonerations (e.g., Miguel Solorio, Abel Soto, and Jofama Coleman) based in part on the new scientific consensus statements that state the importance of focusing on that first test as that first test often contains evidence of innocence (a lineup rejection) instead of evidence of guilt (selecting the suspect's face). Oftentimes, however, that first test is ignored and the same witnesses who provide evidence of innocence on the first test provide misleading "evidence" of guilt at trial by identifying the innocent suspect with

higher degrees of confidence (Garrett, 2011). It is likely that there are many more wrongfully convicted prisoners who are yet to be exonerated using this scientific understanding. The work in this dissertation was designed to enhance our understanding of the often ignored or frequently misunderstood evidence of innocence obtained from lineup rejections.

REFERENCES FOR THE INTRODUCTION AND CONCLUSION

Albright, T.D., Rakoff, J.S., (2020). A clearer view: The impact of the National Academy of Sciences report on eyewitness identification.

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

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, N., Weber, N., & Guerin, N. (2019). Police lineups of the future? *American Psychologist*.

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

Gronlund, S. D., Wixted, J. T. & Mickes, L. (2014). Evaluating eyewitness identification procedures using ROC analysis. *Current Directions in Psychological Science*, 23, 3-10.

National Research Council (2014). Identifying the culprit: Assessing eyewitness identification. Washington, DC: The National Academies Press.

Police Executive Research Forum (2013). A national survey of eyewitness identification procedures in law enforcement agencies.

Wells, G. L., Small, M., Penrod, S. 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*, 3-36.

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

Wixted, J. T., Mickes, L., Clark, S. E., Gronlund, S. D. & Roediger, H. L. (2015). Initial eyewitness confidence reliably predicts eyewitness identification accuracy. *American Psychologist*, 70, 515-526.

Wixted, J. T. & Wells, G. L. (2017). The relationship between eyewitness confidence and identification accuracy: A new synthesis. *Psychological Science in the Public Interest*, 18, 10-65.

Wixted, J. T., Wells, G. L., Loftus, E. F., & Garrett, B. L. (2021). Test a Witness's Memory of a Suspect Only Once. *Psychological Science in the Public Interest*, 22(1_suppl), 1S-18S. https://doi.org/10.1177/15291006211026259

Wixted, J. T., Vul, E., Mickes, L. & Wilson, B. W. (2018). Models of lineup memory. *Cognitive Psychology*, 105, 81-114.

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

Yilmaz, A. S., Shen, K., & Wixted, J. (2024a). The mechanisms of memory and the Federal Rules of Evidence: A psychological perspective. [Manuscript submitted for publication. Preprint available online.] https://doi.org/10.31219/osf.io/hv6bs

Yilmaz, A.S., Wang, X., & Wixted, J.T. (2024b). Response bias modulates the confidence-accuracy relationship for both positive IDs and lineup rejections in a simultaneous lineup task. *Applied Cognitive Psychology*, 38(2), e4196. https://doi.org/10.1002/acp.4196

Yilmaz, A.S., Wilson, B.M., & Wixted, J.T. (2024c). A Rate-them-all lineup procedure increases information and reduces discriminability. *Journal of Experimental Psychology: Applied*. Advance online publication. https://doi.org/10.1037/xap0000524