UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

CRITERIAL VARIABILITY IN EYEWITNESS IDENTIFICATIONS

A THESIS

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

MASTER OF SCIENCE

By

REBECCA LYNN DIDOMENICA
Norman, Oklahoma
2022

CRITERIAL VARIABILITY IN EYEWITNESS IDENTIFICATIONS

A THESIS APPROVED FOR THE
DEPARTMENT OF PSYCHOLOGY

BY THE COMMITTEE CONSISTING OF

Dr. Scott Gronlund, Chair

Dr. Hairong Song

Dr. Adam Feltz

Abstract

Lineups typically induce superior performance compared to one-person identification procedures or showups. Understanding what makes the lineup a superior procedure is important to reduce identification errors, as these can lead to wrongful convictions. Differential filler siphoning, diagnostic feature detection, and criterial variability each attempt to explain lineup superiority. We test the hypothesis that the accuracy of an identification is impacted by group-level variability in criterion placement. Experiment 1 showed greater variability in criterion placement in the showup condition, although overall performance was not worse than the lineup condition. Experiment 2 used different photos of faces from study to test and introduced a constrained showup condition designed to lessen criterial variability by having participants respond only when they are highly confident. Experiment 3 introduced a simultaneous showup condition as another way to help participants set their criterion less variably by making diagnostic information more accessible. We replicate the finding that people set their criterion more variably in showups compared to lineups in Experiments 1 and 2, but this does not translate to a discriminability difference. Experiment 3 results are inconclusive, but this could be due to a small sample size. Therefore, more data are needed to attribute discriminability differences to criterial variability.

Table of Contents

Criterial Variability in Eyewitness Identifications

There are multiple eyewitness identification procedures used by the police. These procedures differentially impact the ability of an eyewitness to discriminate between persons suspected to be innocent or guilty of committing a crime. A witness's decision regarding a perpetrator's innocence or guilt can be the difference between freedom and a correct or erroneous conviction. Coupled with the fact that errors in eyewitness identification decisions are all too common, research that explores the impact of identification procedures on eyewitness decision making is necessitated (Innocence Project, 2021). To further exemplify the importance of this research, consider the following case.

In March 1988, a six-year-old girl was kidnapped and raped near her home. Leonard McSherry was identified by a neighbor as having been loitering in the area where the kidnap occurred and at the time it occurred. To further implicate McSherry in the crime, the four-year-old brother of the girl identified McSherry in a showup. McSherry was sentenced to 48 years in prison for multiple counts of sexual assault. In 1992, new biological evidence surfaced, but McSherry was denied a retrial due to the evidence being insufficient. In 2001, nearly 13 years after McSherry's incarceration, DNA evidence surfaced and proved McSherry's innocence and overturned the conviction.

The injustice here lies, in part, on a reliance on eyewitness identifications as the primary evidence of guilt. Misidentification errors occur at an alarming rate, contributing to the 69% percent of wrongful convictions in the United States (overturned by the Innocence Project). Of particular relevance is the 34% of the 160 exoneration cases, reviewed in 2011, that involved mistaken eyewitness identifications from a police showup (Garrett, 2011). There is a growing body of research suggesting that the initial identification is the most accurate, with mistaken

identifications resulting in lower expressed confidence on average, than accurate identifications (Brewer & Wells, 2006; Sporer, Penrod, Read, & Cutler, 1995; Wixted, Mickes, Clark, Gronlund, & Roediger, 2015; Wixted, Wells, Loftus, & Garrett, 2021). This is problematic for in-court identifications where there is no thorough record of the eyewitness' prior exposure to the suspect, and this threatens the impartiality of these police identification procedures that are used to obtain these identifications.

Despite the body of evidence that different procedures produce different conditions that affect an eyewitness' deliberation of the evidence, there is no outstanding or "optimal" procedure that we can confidently recommend to law enforcement. The "optimal" procedure is the one that maximizes the eyewitness' ability to distinguish between innocent and guilty suspects, i.e., improves *discriminability* (Gronlund, Wixted, & Mickes, 2014). This problem is exacerbated by the fact that different measures reach different conclusions (Gronlund et al., 2014; Gronlund, Carlson, Dailey, & Goodsell, 2009; Lindsay & Wells, 1985). One such measure, the diagnosticity ratio, confounds discriminability with an eyewitness' willingness to choose and is an unreliable measure of performance (Gronlund, Neuschatz, Goodsell, Wetmore, Wooten, & Graham, 2012; Gronlund et al., 2014). The identification procedure with a higher diagnosticity ratio may indicate that the procedure produces more accurate identification performance, induces a greater willingness to choose, or both (Gronlund et al., 2012).

Showups are one-person identification procedures that are frequently used by the police (Smith & Bertrand, 2014; Steblay, Dysart, Fulero, & Lindsay, 2003). To put this into perspective, showups were used in over 50% of the 488 cases (250 cases) recorded in a large U.S. metropolitan area between 1991 and 1995 (Steblay et al., 2003). Showups may be conducted by the police on-sight following a crime thus reducing the likelihood of subsequent

exposure more so than other police identification procedures (Smith & Bertrand, 2014). This bodes well, given the fact that identifications made earlier in time are more reliable and less subject to memory contamination. Timing aside, this procedure is not optimal for distinguishing between guilty and innocent lineup members compared to other lineup procedures used by the police, such as the simultaneous lineup.

Simultaneous lineups are thought to better allow eyewitnesses to discriminate between innocent and guilty lineup members (Colloff & Wixted, 2020; McAdoo & Gronlund, 2016; Smith, Wells, Lindsay, & Penrod, 2017; Wells, Smalarz, & Smith, 2015; Wetmore, Gronlund, Neuschatz & McAdoo, 2017; Wixted & Mickes, 2014). Researchers refer to this difference in discriminability as the "lineup advantage". We focus on simultaneous lineups (all lineups presented at once) in our research because recent research suggests that simultaneous lineups are diagnostically superior to sequential lineups, which involves lineup members being presented in succession (Amendola & Wixted, 2015; McQuiston-Surrett, Malpass, & Tredoux, 2006; Steblay, Dysart, & Wells, 2011; Wixted, Mickes, Dunn, Clark & Wells, 2016). This is due, in part, because suspect identifications are more diagnostic of guilt in simultaneous lineups and the accuracy of these identifications are not conditional on other factors being present, such as in the sequential lineup (Carlson, Gronlund, & Clark, 2008; Wixted et al., 2016).

Understanding what makes the simultaneous lineup a superior procedure is important to reduce the risk of future injustices, but it also is of theoretical interest. If simultaneous lineups are to be recommended to the police over showups, the underlying theoretical rationale should be a key component of that recommendation (McQuiston-Surrett, Malpass, & Tredoux, 2006). Showups and simultaneous lineups may recruit different cognitive processes, highlighting the

need for an evaluative comparison of these procedures (Steblay, Dysart, Fulero & Linsdsay, 2003).

Currently, the lineup advantage is explained by contextual variables, such as lineup and showup compositional differences. For example, the showup presents one face to the witness for identification, whereas the lineup presents up to six faces simultaneously. Presumably, the procedure with more faces offers up more information, and this information aids identification decisions (Wixted & Mickes, 2014). The above definition of "optimal" could be expanded to include the procedure that best preserves the eyewitness evidence. We believe these composition-centered explanations are incomplete; the presence of information alone does not dictate how the eyewitness uses the information. The eyewitness could require more or less information to make their decision, counterintuitively discarding some of the information as non-diagnostic, or be inconsistent with the amount of information they require to make an identification across procedures. Each of these actions could assess the evidence differently, or discount the evidence altogether in a way that would threaten the optimality of decision making in this context.

We begin with an overview of the identification procedures. Then, we review the different explanations for the lineup advantage. This includes one explanation that we favor, in which the variability in the amount of evidence needed to separate innocent from guilty suspects across eyewitnesses, the criterial variability, is greater for showup identifications, which negatively impact these identifications compared to lineup identifications.

Identification Procedures

In a showup, either a guilty suspect or an innocent suspect is shown to the witness. An identification is either made, or the face is rejected. Target absent showups contain the innocent

suspect whereas target present showups contain the guilty suspect. An identification is either made, or the showup is rejected (no identification is made). Identifications are correct for target present showups, but incorrect for target absent showups whereas non-identifications are correct for target absent showups, but incorrect for target present showups.

A simultaneous lineup is a procedure where faces, typically six, are presented to the witness all at once. Faces selected for the lineup commonly are matched to the description of the perpetrator provided by the eyewitness (Colloff &Wixted, 2019; Luus & Wells, 1991). In a fair simultaneous lineup, the fillers resemble the suspect (innocent or guilty). Target absent lineups contain only fillers whereas target present lineups contain both fillers and the guilty suspect. An identification is either made, or the lineup, including all faces, is rejected. Identifications are correct for target present lineups in which the selected face is the guilty suspect, but incorrect for target absent lineups. The rejection of the lineup is correct for target absent lineups, but incorrect for target present lineups.

Predicting overall performance in these tasks depends on whether the guilty suspect is present (target present) or replaced with a designated innocent suspect or filler (target absent). Signal detection theory (SDT) is used to interpret how a witness who makes correct identifications over many successive tests of their memory is, on average, an accurate witness, and how a witness who makes incorrect identifications over many successive tests of their memory is, on average, an inaccurate witness. The procedure that promotes greater accuracy in responding is the superior procedure (ignoring other factors). We turn next to signal detection theory to explain how our data will be analyzed.

Signal Detection Theory (SDT)

SDT is a framework for understanding how memory evidence is handled at the time of a decision (e.g., Green& Swets, 1966; Kellen, Klauer & Singmann, 2012). Evidence for familiar or previously encountered stimuli (new) and unfamiliar or not previously encountered stimuli (old) accumulates to some degree of strength that determines the type of decision that is made. Lower memory strengths are represented by lower values along the memory strength axis and represent less accumulated memory evidence overall (Figure 1).

We can use SDT to answer questions about eyewitness reliability because it summarizes performance in terms of discriminating between seen and unseen stimuli and these discriminations approximate the old/new decisions that eyewitnesses make. In eyewitness identifications, a familiar object i.e., the perpetrator, has a higher associated memory strength than an unfamiliar object, i.e., the innocent person. The memory strengths associated with the perpetrator are assumed to be normally and continuously distributed and are summarized by a probability distribution (see Figure 1). The distribution on innocent faces is positioned to the left of the distribution on perpetrator faces along the memory strength axis in Figure 1. These distributions are hereafter referred to as the filler and target distributions, respectively. Any witnessed details of a crime can map onto these distributions, including memory for the suspect's face. For example, if a witness gets a good, long look at the perpetrator, there is more evidence on which to base their decision, resulting in greater separation between the filler and target distributions. Conversely, if a witness only gets a short glance at the perpetrator, there is less memory evidence on which to base a decision, resulting in more overlap between the filler and the target distributions (because the position of the target distribution shifts down the memory strength axis).

Memory for innocent and suspect faces can be summarized by two signal detection-based measures, *discriminability* and *response bias*. Discriminability reflects how well guilty suspects can be distinguished from innocent suspects. In SDT, discriminability is a function of the distance between the mean of the guilty suspect or target distribution and the mean of the filler distribution, scaled against the variability of the distributions, and is indicated by a measure called d prime (*d'*)[1]. Higher values indicate a stronger ability to discriminate between innocent and guilty suspects compared to chance performance (*d'* = 0).

The second component of recognition performance is the response bias and is indicated by the measure, *c*[2]. In eyewitness identification experiments, this is the participant's overall willingness to identify a face as suspect or "familiar". Response bias is captured by the placement of the decision criterion, represented by the vertical line in Figure 1 (response bias = 0). Placement of the decision criterion is generally assumed to be under the control of the witness who can decide what degree of memory strength separates "familiar" from "unfamiliar" decisions (Benjamin, Diaz, & Wee, 2009; Mueller & Weidemann, 2008). Faces that exceed the decision criterion and are deemed "familiar" and result in either a hit or a false alarm, depending on the trial type.

Next, we will look at the explanations offered for the lineup advantage that do not rely on the assumptions of signal detection theory but are still adequate for explaining the advantage.

Diagnostic Feature Detection Theory

According to the diagnostic feature detection theory, diagnostic features can be used to distinguish between innocent and guilty suspects in a fair lineup where all fillers match the

[1] *d'* = z(HR) - z(FAR), where z() is the inverse of the cumulative distribution function of the underlying distribution.
[2] *c* = –½ [z(HR) + z(FAR)]

suspect on general characteristics (Wixted & Mickes, 2014). The memory signal associated with a face is a composite of facial features and other identifiable aspects such as race, and age. The witness identifies features common among faces in the lineup as non-diagnostic, and those that are unique to the perpetrator as diagnostic, and discards those non-diagnostic features from the identification decision (Wixted & Mickes, 2014).

Discriminability should increase with the availability of these diagnostic cues (Wixted & Mickes, 2014). The greater the diagnostic cues, the greater the ability to discriminate between fillers and guilty suspects and the lesser the degree of overlap between their respective evidence distributions. There are invariably more diagnostic cues available in a lineup compared to a showup because only one face is viewed in a showup, making the process of distinguishing diagnostic from non-diagnostic cues challenging. Therefore, diagnostic feature detection theory predicts that discriminability should be worse in a showup.

This explanation is sufficient to capture the magnitude of the lineup-showup difference seen in empirical data, however, it does not make any predictions for criterial variability. It is possible that criterial variability is interacting with cue accessibility to improve discriminability. As a possible interaction, the more diagnostic cues there are available, the more willing the participant is to identify a face independent of their prior willingness to choose. By this account, diagnostic feature detection is a plausible explanation, but one that does not attempt to parse decision noise from memory noise.

Differential Filler Siphoning

Differential filler siphoning theory posits that the fillers in a fair target-absent lineup protect an innocent suspect from being chosen by siphoning choices away from the innocent suspect to the fillers because they compete for choices (Smith et al., 2017; Wells et al., 2015).

This siphoning effect is seen as advantageous, as it reduces the rate of misidentifications (i.e., false alarms) (Colloff & Wixted, 2020). In target present lineups, because targets are, on average, more memorable than fillers, siphoning occurs disproportionately less compared to target-absent lineups thus making this process differential (Smith et al., 2017; Colloff & Wixted, 2020).

Evidence from simulations of the WITNESS model (Clark, 2003), a direct-access matching model designed to capture aspects of the eyewitness decision process, reveals that the presence of fillers in a lineup may not induce a large enough lineup advantage to match empirical data (see Wetmore et al., 2017). Therefore, filler siphoning theory may be less sufficient to explain the lineup advantage than previously determined. Furthermore, this explanation has limited predictive ability as it only makes predictions regarding discriminability, ignoring other components of the decision process such as decision noise.

As an alternative to the diagnostic feature hypothesis, and as a potential supplemental mechanism to add to filler siphoning, Wetmore et al. (2017) proposed that the magnitude of the empirical lineup advantage can be better approximated if decision noise (criterial variability) is added to the model (Wetmore et al., 2017; Benjamin et al., 2009). Before we can determine whether criterial variability acts as a supplemental mechanism to filler siphoning or diagnostic feature detection, we must understand and solidify its role in recognition memory performance.

The Criterial Variability Hypothesis

Setting one's criterion is an inherently noisy process (Benjamin, Diaz, & Wee, 2009; Mueller & Weidemann, 2008). To illustrate this point, some eyewitnesses may choose to adopt a conservative criterion making them less likely to endorse a face as having previously been seen unless they are confident, whereas other eyewitnesses witnessing crime may choose to adopt a liberal criterion making them more likely to endorse a face as having previously been seen,

regardless of their confidence level. This variability in criterion placement across eyewitnesses is a source of decision "noise" that increases the overlap between the evidence distributions, thereby decreasing discriminability (see Benjamin, Diaz, & Wee, 2009).

Criterial variability is estimated by the standard deviation of the average criterion placement across participants. To obtain estimates of discriminability and criterion placement, we use the hit and false alarm rates estimated from each participant. Conceptually, the hit rate is the proportion of the target distribution that falls above the decision criterion ($c = 0$ in this example), whereas the false alarm rate is the proportion of the filler distribution that falls above the criterion (Figure 1). The false alarm rate decreases if a more conservative criterion is adopted ($c < 0$ in this example), as the proportion of the filler distribution that falls above the criterion reduces relative to its position indicated in Figure 1. Conversely, the hit rate decreases if a more conservative criterion is adopted, and increases if a more liberal criterion is adopted. A higher *d*' occurs when the hit rate exceeds the false alarm rate (there is greater endorsement of old decisions for old versus new trials). Chance performance occurs where the hit rate and false alarm rates are equal (there is an equal endorsement of old decisions for old and new trials). The false alarm rate increases when a more liberal criterion is adopted ($c > 0$ in this example), as the proportion of the filler distribution that falls above the criterion increases relative to its position indicated in Figure 1. We capitalize on the fact that criterion placement can be easily manipulated in our test for this hypothesis.

Predictions

Whereas diagnostic feature detection and filler siphoning do not make any predictions regarding criterial variability, the criterial variability hypothesis predicts that criterial variability contributes to eyewitness decisions (Wetmore et al., 2017). The criterial variability hypothesis

proposes that showup witnesses set their decision criterion more variably compared to lineup witnesses, and this plays a role in reducing discriminability (Wetmore et al., 2017; see Experiment 1).

In a showup, participants cannot use other faces to calibrate where they want to place their criterion, which results in greater variability in placement across participants. In a lineup, on the other hand, the other faces help participants better calibrate where to place their decision criterion, resulting in less criterial variability. If showup witnesses have greater difficulty setting their criterion compared to lineup witnesses it supports the idea that decision noise plays a role in discriminability as summarized in the criterial variability hypothesis (see McAdoo & Gronlund, 2016; Wetmore et al., 2017; Starns, 2014). Therefore, we seek empirical evidence for greater criterial variability in the showup condition as a factor contributing to reduced showup discriminability.

We hypothesize that there is an adverse impact of criterial variability on discriminability such that the greater the criterial variability, the more overlap between distributions, and the worse the discriminability. *If we find evidence that participants are setting their criteria more variably in showups compared to lineups, and this is accompanied by a discriminability decrease, then criterial variability is a source of decision noise that negatively impacts showup decisions.*

Evaluating the differential filler siphoning and diagnostic feature detection explanations alongside the criterial variability hypothesis will help determine the extent to which decision noise is at play. We now transition to the current state of the research and what we propose to be the next steps in evaluating criterial variability as a viable explanation for the lineup advantage.

Experiment 1[3]

Experiment 1 was designed to test the idea that the lineup advantage occurs, at least in part, due to showup witnesses setting their criterion more variably than lineup witnesses (McAdoo & Gronlund, 2016; Wetmore et al., 2017).

Method

Participants

Participants were enrolled in introductory psychology courses at the University of Oklahoma and completed the study in exchange for partial course credit. Data from 205 participants (White or Caucasian (76.21%), American Indian or Alaska Native (4.86%), Asian (8.25%), Black or African American (6.31%), Native Hawaiian or Pacific Islander (0.49%), Middle Eastern (1.46%), and No Response (2.43%)) was retained for analysis. There were 156 females and 49 males and the average age was 18.60 years old (*SD* =1.95 years).

Materials

Faces were sampled from the Bainbridge et al. (2013) database and were Caucasian males between the ages of 18-45. Three Caucasian female celebrities were sampled from this database and included in the practice phase.

Procedure

This experiment included a practice phase, a study phase, and a test phase. In the practice phase, participants completed the task for three Caucasian female celebrities from the Bainbridge et al. database. In the study phase, 50 target faces were shown for 3 seconds each, with a 500 millisecond inter-stimulus fixation. In the test phase, target present trials included one target and no fillers in the showup condition, or four fillers in the lineup condition. Target absent trials

[3] Ryan McAdoo, Ph.D., and Kylie Key, Ph.D. designed, executed, and analyzed this experiment.

included no targets, and either one filler in the showup condition, or five fillers in the lineup condition. In the showup condition, participants used keys 1 and 2 to indicate if a face had previously been seen (1= "yes") or had not previously been seen (2 = "no"). In the lineup condition, participants used keys 1-5 to indicate the position of the face that they believed to have previously seen using a response key, or indicated "0" for "not present". Then, participants expressed confidence in their decision on a scale of 1 (not at all confident) to 9 (extremely confident). Lineup versus showup was manipulated between-subjects.

Results

Discriminability and criterion placement was computed for both conditions. Summary statistics are available in Table 1. An independent samples *t*-test revealed no significant mean difference in *d*' between showups and lineups, $t(146) = -.21$, $p = .83$. However, there was a significant difference in criterion placement (denoted *c*), with lineups having the more conservative criterion placement, on average, compared to showups; $t(146) = -13.91$, $p < .001$. Most importantly, there was a significant difference in criterion variability, with Levene's test statistic = 25.27, $p < .001$, supporting the hypothesis that showups have greater variability in criterion placement than lineups.

Discussion

Experiment 1 offers moderate support for the hypothesis. Participants in the showup condition placed their criteria more variably, on average, compared to participants in the lineup condition. However, no discriminability differences were detected. The lack of a discriminability difference replicates another experiment that included multiple study-test trials. Meissner et al. (2005) compared the effect of lineup size on discrimination accuracy and criterion placement. Criterion placement became more conservative lineup size increased. However, endorsing more

conservative criteria in the lineup condition did not improve overall discrimination accuracy in this condition. Rather, discrimination accuracy was better in showups where liberal criteria were endorsed, and this pattern of results held regardless of the lineup size (Experiment 3, Table 3, Meissner et al., 2005).

Differences in criterial variability were significant and in the predicted direction, but this was not accompanied by a discriminability difference. However, the fact that we found support for one of our two expected findings is encouraging. Perhaps criterial variability is negatively impacting showup decisions but not in the way we had initially proposed. To explore these discriminability and criterial variability differences further, we designed Experiment 2.

Experiment 2

If criterial variability is a source of decision noise that harms performance, if it is reduced, this should have the impact of improving discriminability. In Experiment 2, we included instructions in a showup constrained condition (SU_c) designed to help participants set their criterion less variably. If our manipulations to constrain criterion placement are effective, and participants place their criterion over a smaller range, they should also be setting a less variable criterion. Support for the criterial variability hypothesis would be indicated if reduced variability in the constrained showup condition improves discriminability.

We predicted that the range of possible criterion placements along the strength of evidence axis for participants should be the greatest in the showup condition, followed by the constrained showup, and then the lineup conditions (see Figure 2). Notice in Figure 2 that the same average criterion position is maintained despite changes to the range. This is for illustrative purposes. If instructions were to result in a different average criterion (criterion shift), this would not invalidate the hypothesis.

Method

Participants

Participants were recruited from the research participation system (SONA) at the University of Oklahoma and through a mass email sent to all OU undergraduates. Participants recruited through the research participation system were enrolled in introductory psychology courses at the university and were 18 years of age or less than 18 years of age and compliant with SONA. Data from 72 participants (56.9% White, 18.1% Asian, 11.1% Black or African American, 6.9% American Indian or Alaskan Native, 1.4% Middle Eastern, and 5.6% no response; of these non-responses, half indicated that they were of Hispanic or Latino origin) was retained for analysis. There were 38 females and 34 males and the average age was 19.96 years old (*SD* =2.67 years).

Materials

Faces were sampled from the SCface database and were primarily Caucasian (Grgic et al., 2011). The SCface Database offers a collection of faces that were taken at different points in time for the same individual (Grgic, Delac, & Grgic, 2011). This new, larger, database allowed us to use different faces at study and test (an improvement in ecological validity), and to make fairer lineups than what were previously constructed for Experiment 1. Faces were selected if they met the minimal criteria for selection, that being that the face appeared to be between 25 and 40 years old. We did not make an attempt to counterbalance males and females, but we did ensure that both were present. A practice phase included three celebrity faces from the Bainbridge et al. (2013) database.

100 faces were selected from the Grgic et al. (2011) database. Of these 100, 50 were randomly assigned as targets. To construct target present lineups, four fillers were assigned to a

target if they matched on the following general characteristics: hair color, hair length, race, age, and gender. Target absent lineups included five fillers that were selected based on the above characteristics. Once the faces from the SCface database were exhausted, additional fillers were selected from statewide, publicly available criminal databases for Florida and Ohio. These faces were selected based on the researcher's discretion to resemble existing SCface database fillers until all lineups were constructed. We note that participants were not tested on the same set of target absent stimuli across conditions, despite being tested on the same set of target present stimuli across conditions. This error confounds the stimuli with the condition and threatens the internal validity of the study as well as our ability to detect true differences between conditions on account of the treatment. This problem was corrected in Experiment 3.

The experiment was constructed in E-Prime 3.0 and converted into a distributable E-Prime Go file for remote data collection. Study materials, procedure, and guidelines were approved by the University of Oklahoma IRB and followed APA ethical guidelines.

Procedure

Participants were randomly assigned into to one of three conditions; the showup condition (SU), the constrained-showup condition (SU_c), or the lineup condition (LU). As this was an online study, research assistants emailed out the link to a Zoom meeting, a link to the experiment, and the session number and subject ID. In the Zoom meeting, participants downloaded the link to the experiment and entered in their assigned subject ID and demographic information into the starting prompts. Research assistants monitored participants as they read the consent form and pre-study instructions, and signed off of the Zoom meeting following completion of the pre-study instructions.

In all conditions, participants were instructed that they would be presented a series of faces, one after the other, and were instructed to memorize these faces using no particular strategy. They were also instructed that they would be tested on these faces later. Participants then completed a practice phase including two target present trials and one target absent trial (in no particular order). In the study phase, a total of 50 faces were shown for three seconds each with a 500 millisecond inter-stimulus fixation. Following the study phase, participants were presented with a distractor task wherein they indicated whether two numbers summed to the third number using "y" (yes) and "n" (no) keys.

In the test phase of the showup and constrained-showup conditions, a total of 50 target present and 50 target absent trials were randomly ordered. In the test phase of the lineup condition, 50 target present and 50 target absent 5-person lineups were randomly presented to participants. Participants were instructed that not all lineups or showups that they see may contain the guilty suspect. To reinforce this, participants used keys 1-5 to indicate the position of the face that they believed to have previously seen using a response key as a guide (see Figure 3), or indicated "0" for "not present". Then, participants expressed confidence in their decision on a scale of 1 (not at all confident) to 9 (extremely confident).

In the constrained showup condition, participants were advised that the choosing rates for innocent suspects in the population is high, and that they should only choose a face from the showup if they are highly confident. Participants used keys 1 and 2 to indicate if a face had previously been seen (1= "yes") or had not previously been seen (2 = "No"). Responses were self-paced. Following the test phase, participants were debriefed via email and were either compensated with course credit (SONA) or a $10 Amazon gift card[4] (mass email).

[4] Gift cards were provided by the Department of Psychology at the University of Oklahoma.

Results

All preliminary analyses were conducted after collapsing across conditions. Shapiro-Wilk, Anderson-Darling, and Cramer-von Mises tests for normality suggest that *d'* and *c* measures are normally distributed (see Appendix). Tests for homogeneity of variances suggest that the group variances for *d'* are equal, whereas the group variances for *c* are unequal (see Appendix). Residual plots reveal systematic differences with respect to the fitted model for *d'*, but not *c* (see Appendix). Neither trial-level nor participant-level exclusions were performed. Response times for the identifications did not fall outside the range of reasonable response times across participants (beginning at 300 milliseconds) (Ratcliff, 2006; Ratcliff & Starns, 2009).

A one-way ANOVA was conducted to examine whether discriminability differed according to the type of identification procedure used. Discriminability was not significantly different among conditions; $F(2,69) = 1.73$, $p = .19$ (see Table 2).

A one-way ANOVA was conducted to examine whether criterion placement differed according to the type of identification procedure used. Criterion placement was significantly different among conditions; $F(2,69) = 13.64$, $p < .05$, $\eta^2 = 0.28$ (see Table 2). Pairwise comparisons were tested post hoc, and a Bonferroni correction was applied. Participants responded most conservatively in the lineup condition, and most liberally in the showup condition ($p < .01$, 95% *CI* of the mean difference = 0.430 to 0.847). The average criterion placement in the constrained showup condition was significantly greater than average criterion placement in the showup condition ($p < .01$, 95% *CI* of the mean difference = -0.616 to -0.187), but not the lineup condition ($p = .18$).

Levene's test for equality of variances was used to assess between group differences in criterial variability. Criterial variability was significantly different among conditions, Levene's

test statistic = 5.96, $p < .05$. Pairwise comparisons were tested post hoc, and a Bonferroni correction was applied in a pairwise fashion. The measure of effect size for Levene's test was also computed in a pairwise fashion and is the log transform of the ratio of the coefficient of variation between two groups. This measure is deemed useful for between group comparisons on variability (Nakagawa et al., 2015). Critically, the average variability in criterion placement was the least variable in the lineup condition, and significantly different than the average criterial variability in showups ($p < .05$, $\log (CV_{LU} / CV_{SU}) = 0.237$) but not the constrained showup ($p = .08$, $\log(CV_{LU} / CV_{SU}) = 0.316$). Of primary interest, the average criterial variability was significantly greater in constrained showups than in showups ($p < .05$, $\log (CV_{SU} / CV_{SUc}) = 0.871$).

Discussion

The results offer little support for the hypothesis. Our attempts to reduce variability in criterion placement were not effective in that the variability in criterion placement in the constrained showup condition was not reduced relative to showups. Variability in criterion placement in the lineup condition was smaller than in the showup condition, however, this did not translate to a significant discriminability difference between these conditions. Therefore, results do not support our claim that the lineup advantage occurs, at least in part, because participants set a more consistent criterion in the constrained showups compared to showups. Furthermore, with no differences in discriminability, we did not replicate the lineup advantage. However, we interpret these results with caution given the aforementioned confounding. It is possible that the target absent stimulus differences are driving these effects and not the manipulation itself.

Perhaps more statistical power is needed to detect discriminability differences, and to then attribute these differences to criterial variability. Specifically, more data may be needed to determine if setting a more consistent criterion translates to better performance. We designed an Experiment 3 to test the criterial variability hypothesis under different conditions, eliminating the confounding, and using an approach that we believed could more effectively constrain criterion placement.

Experiment 3

We designed Experiment 3 as another attempt to reduce criterial variability and enhance discriminability. We included a simultaneous lineup condition (SU_{sim}) in place of the constrained showup condition and tested this condition alongside the lineup (LU) and showup (SU) conditions. Given that our attempts to constrain criterion placement through showup instructions in the constrained showup condition were not effective, we thought that this approach could more effectively constrain criterion placement. A simultaneous showup is a type of identification procedure that, like a lineup, presents similar looking faces alongside a suspect. However, the witness can only identify the suspect and not the other faces (Colloff & Wixted, 2019). This procedure has been used to evaluate the diagnostic feature detection theory, but not the criterial variability hypothesis.

We expect the simultaneous showup to behave like a lineup. Seeing other faces in a simultaneous showup should help participants set their criteria less variably. For our hypothesis to be supported, we expect that criterial variability is reduced in this condition relative to showups, and this improves discriminability relative to showups. We do not have any specific predictions for the magnitude of criterial variability and discriminability differences between the

simultaneous showup and lineup, but non-significant differences are plausible and would support the hypothesis.

Method

Participants

A power analysis was conducted in GPower using the default alpha level (0.05) and Cohen's *f* effect size (0.25). Power was set to 0.8 and a one-way ANOVA was specified. This analysis revealed that 159 participants (53 participants per condition) are needed to detect effects in the data and achieve adequate power. Despite persistent efforts to recruit a sufficient sample size, by our deadline, only 53 participants were recruited from the research participation system (SONA) at the University of Oklahoma (69.8% White, 13.2 % Asian, 11.3% Black or African American, 3.8% American Indian or Alaskan Native, and 1.9% no response. Four participants (7.5%) indicated that they were of Hispanic or Latino origin in addition to their indicated race). There were 40 females and 13 males, and the average age was 19.48 years old (*SD* = 4.12 years). The achieved power across samples was 0.32, resulting in a small to medium effect size.

Participants were enrolled in introductory psychology courses at the university and were 18 years of age or less than 18 years of age and compliant with SONA. Participants were compensated with course credit (up to 1 credit total).

Materials

Lineup stimuli were carried over from Experiment 2. We evaluated the fairness of these lineups by checking to see if there were any lineups that incurred a 50% rejection rate across participants *and* had at least one filler that did not get chosen (target present lineups), and only three, two or one fillers that got chosen (target absent lineups). These fillers were replaced with fillers that more closely matched the suspect on hair color, hair length, race, age, and gender.

Of those remaining lineups, and using the Experiment 2 data, we identified 10 target present and 10 target absent lineups that incurred the least number of correct responses across participants and eliminated these lineups (see code in Appendix). Consequently, we tested fewer lineups than were in Experiment 2. All things considered, of the 100 original lineups, 70 were retained.

The experiment was constructed in E-Prime 3.0 and converted into a distributable E-Prime Go file for remote data collection[5]. Study materials, procedure, and guidelines were approved by the University of Oklahoma IRB and followed APA ethical guidelines.

Assessing Lineup Fairness

We assessed the fairness of ten target present lineups that were randomly sampled from the pool of 70 remaining lineups. The designated target in each lineup was placed into a Qualtrics survey for evaluation. Ten research assistants were recruited for this task. Research assistants were asked to view each face and provide a description and explicitly mention age, hair color, hair style, rage, and gender. Only the most popular descriptors were used to create a modal description for each target. Then, a separate Qualtrics survey was administered to 36 participants completing the survey for course credit in SONA using these modal descriptions. These participants selected the face in the lineup (using keys 1-5) that they believed best matched the modal description for that lineup. Tredoux's effective size (*E*') was computed for each lineup given the pattern of responses generated from this survey and using the 'r4lineups' package in R (Tredoux, 2018). The average effective size across lineups was 2.1 indicating that

[5] Seven participants completed the experiment in-person.

there were, on average, about two plausible lineup members, making these lineups moderately unfair (see Table 3)[6].

Procedure

Participants were randomly assigned into one of three conditions; the showup condition (SU), the lineup condition (LU), and the simultaneous showup condition (SU_{sim}). As this was an online study, research assistants provided the session number and subject ID to participants over email prior to the session. Participants clicked on the link and entered their assigned subject ID and demographic information in the starting prompts to initiate the experiment. Participants then reviewed the consent form and pre-study instructions. These instructions informed participants that they would be presented a series of faces, one after the other, and could memorize these faces using no particular strategy. They were also instructed that they would be tested on these faces later.

Each condition included two blocks with a study phase, distractor task, and a test phase. In the study phase, in the first block, participants studied 20 faces for 4 seconds each followed by a 500-millisecond fixation. These faces were randomly ordered. Following the study phase, participants were presented with a distractor task that required them to indicate whether two numbers summed to the third number using "y" (yes) and "n" (no) keys.

In the test phase, 20 target present and 20 target absent trials were randomly ordered. All responses were self-paced. A single photo was presented to participants in the SU condition.

[6] The suspect is chosen by participants at a rate that is greater than chance (20%) in all cases (see Appendix). However, because the effective size (2.1) is lower than the nominal lineup size (5), it is inaccurate to say that these lineups are biased (Malpass, Tredoux, & McQuisition, 2007). Rather, if a witness randomly chooses among the two plausible lineup members, the risk of mistaken identification increases from 20% to 50%. Therefore, lineups with suspect choosing rates that exceed 50% are biased.

Participants used keys 1 and 2 to indicate if a face had previously been seen (1= "yes") or had not previously been seen (2 = "no"). 5-photo lineups were presented to participants in the LU condition, and participants used keys 1-5 to indicate the position of the face that they believed to have previously seen using a response key as a guide (see Figure 3), or indicated "0" for "not present". 5-photo lineups were presented to participants in SU_{sim} (modified from Colloff and Wixted, 2019). In this condition, the suspect (guilty in target present trials, innocent in target absent trials) was outlined in red (see Figure 4). Participants were told that they could only identify the face outlined in red. Participants were also told that the non-outlined faces were included to help them decide whether they have studied the face outlined in red before or not, but that these faces were *not* previously studied. Instructions prompted participants to indicate if the face outlined in red had previously been seen (1 = "yes") or had not previously been seen (2 = "no"). In all conditions, participants expressed confidence in their decision on a scale of 1 (not at all confident) to 9 (extremely confident).

Participants were allotted a break in between blocks. The procedure was repeated in the second block using the remaining set of 20 target present and target absent lineups and showups. The stimuli were randomized within blocks, but not between blocks for each subject. That is, all subjects viewed the same 20 faces in block 1, and the same 20 faces in block 2, though these faces appeared in a random order for each subject. Following the final test phase, participants were debriefed and received a half credit in SONA (up to 1 credit total).

Results

All preliminary analyses were conducted after collapsing across conditions. Shapiro-Wilk, Anderson-Darling, and Cramer-von Mises tests for normality suggest that *d'* and *c* measures are normally distributed (see Appendix). Tests for homogeneity of variances suggest

that the group variances for *d'* and *c* are equal (see Appendix). Residual plots reveal systematic differences with respect to the fitted model for *d'* and *c*, supporting the homogeneity of variances tests (see Appendix). Neither-trial level nor participant-level exclusions were performed. Response times for the identifications did not fall outside the range of reasonable response times across participants (beginning at 300 milliseconds) (Ratcliff, 2006; Ratcliff & Starns, 2009).

A one-way ANOVA was conducted to examine whether discriminability differed according to the type of identification procedure used. Discriminability was not significantly different among conditions; $F(2,50) = 0.30$, $p = .75$ (see Table 4).

A one-way ANOVA was conducted to examine whether criterion placement differed according to the type of identification procedure used. Criterion placement was significantly different among conditions; $F(2,50) = 19.40$, $p < .05$, $\eta^2 = 0.44$ (see Table 4). Pairwise comparisons were tested post hoc, and a Bonferroni correction was applied. Participants responded most conservatively in the lineup condition, and most liberally in the showup condition ($p < .01$, 95% *CI* of the mean difference = 0.360 to 1.098). The average criterion placement in the simultaneous showup condition was significantly smaller (more liberal) than the average criterion placement in the lineup condition ($p < .01$, 95% *CI* of the mean difference = 0.445 to 0.716). The average criterion placement was not significantly different between the simultaneous showup and showup conditions ($p = .96$). Of primary interest, criterial variability was not significantly different among conditions, Levene's test statistic = 0.174, $p = .84$ (see Table 4).

Discussion

Criterial variability was reduced in the simultaneous showup condition relative to the showup condition, however, this difference was not significant and did not impact

discriminability in the predicted direction. We also proposed that participants would set their criterion more variably in showups compared to lineups. Perhaps these differences will reach significance after data collection is complete. We did not replicate the lineup advantage; no discriminability differences were detected. Criterion placement did vary significantly according to the type of identification procedure used, but this result alone does not allow us to test our predictions.

General Discussion

The present research explored the impact of average variability in criterion placement across identification procedures on identification accuracy. We were interested in seeing whether criterial variability directly impacted discriminability, and whether this relationship depended on the type of identification procedure used. Specifically, we were interested in seeing whether reducing criterial variability (through showup instructions or the addition of photos for context) could improve discriminability by reducing criterial variability.

We began with the premise that because each procedure has a different composition, each procedure could possibly warrant a different approach to probing the memory evidence. That is, participants could set their criteria to evaluate this evidence in a more or less consistent manner across trials, depending on what was warranted by the procedure. For instance, in the constrained showup condition, participants may choose to adopt a more consistent criteira across trials because they approach the task with the mindset to reduce identification errors. In the simultaneous showup condition, seeing other faces should help participants refine their diagnostic cues and set a more consistent criteira (thus reducing criterial variability overall).

In Experiment 1, we were able to detect a significant difference in criterion placement between the lineup and showup conditions, with participants placing their criteria more variably

in the showup condition than in the lineup condition. However, no discriminability differences were detected, so we were unable to replicate the lineup advantage. In Experiment 2, we tested participants using a more ecologically valid set of stimuli and included a constrained showup condition. Criterial variability was not reduced in the constrained showup condition relative to showups, so our manipulation to constrain criterion placement in this condition was not effective. Criterial variability in the lineup condition was significantly less than in the showup condition, however, this did not translate to a significant discriminability difference. Without a discriminability difference, we were unable to replicate the lineup advantage. In Experiment 3, we increased the encoding time from three to four seconds and included a simultaneous showup condition. We had no significant findings in terms of our key measures; the results for criterion placement, while significant, do not offer support for the criterial variability hypothesis.

The lack of support for the hypothesis in Experiment 2 could be explained by the confounding in this experiment. By testing different target absent stimuli in different conditions, we are unable to separate the effect from the target absent stimuli from the manipulation itself. The lack of significant findings in Experiment 3 may be explained by a small sample size. Our achieved power was only 0.32, resulting in an ability to detect only a small to medium effect size.

Overall, these results offer little support for our hypothesis and do not lend themselves to any practical application, such as recommending the lineup over the showup to law enforcement. We saw differences in criterial variability in the predicted direction in two out of the three experiments, but this did not translate to improvements in discriminability. We were unable to replicate the lineup advantage in any of the three experiments. These results suggest that simply

reducing variability in criterion placement is not enough to improve performance in these tasks, and overall, does not appear to be the driving mechanism for improving discriminability.

Limitations

A clear limitation in Experiment 2 is the confounding of the target absent stimuli with the conditions. Testing the participants on a different set of target absent trials across conditions threatens our ability to detect true differences between conditions on account of the treatment. Another clear limitation is the small sample size in Experiment 3. Our achieved power was only 0.32, resulting in an ability to detect only a small to medium effect size. We hope to be closer to our sample-size goal by the time that data collection is complete.

The average discrimination in Experiment 2 is poor, albeit above chance ($d' = 0$). Experiment 2 and Experiment 3 were conducted remotely due to the pandemic. It is possible that participants' motivation and overall success contributed to this result. Relatedly, the task may have been too difficult for participants to navigate on their own; we could not provide the same level of assistance online that we could in the lab.

Presenting faces in a single-trial format may allow criterial variability differences to reveal themselves better than in multiple-trial formats, as well as reduce testing and fatigue effects. Meissner et al. (2005) did not detect discriminability differences between lineups and showups in their multiple-trial experiment. In lacking discriminability differences, we cannot demonstrate that criterial variability improves discriminability, which is a central prediction of the criterial variability hypothesis.

To the extent that identifications depend on the base rate of cues (i.e., facial hair, age) in the population, we could have overestimated identification accuracy in lineups where these cues are present. The stimuli were similar in age and appearance to the student population that

provided the data. Participants may have noted the prevalence of, for example, goatees in the population and were thus predisposed to selecting faces with this feature (despite efforts to match on these characteristics).

The measures we use to evaluate performance in these tasks are not theory-free (Brady et al., preprint). To derive these SDT measures, we are assuming that the underlying memory strengths for faces are continuously and normally distributed (Green & Swets, 1966). This is not an unreasonable assumption; however, responses in these tasks are binary and this limits our ability to examine the underlying shapes of the evidence distributions, and this can undermine our ability to capture true performance (Brady et al., preprint). For example, a participant that adopts liberal criteria is more accurate for genuinely new and genuinely old items than a participant who adopts conservative criteria (Brady et al., preprint). However, because they opted for a liberal criterion, their false alarm rate is inflated, and their discriminability is deflated (Brady et al., preprint). Whether the criteria were a true reflection of their memory or not, if the criteria are set consistently across trials, we do have an accurate measure of criterial variability.

Future Directions

If criterial variability is not acting on discriminability to improve performance, what is it acting on, and what is acting on discriminability? Perhaps there is some combination of all three explanations that could account for these results. Research that tests these explanations together to parse out their contributions is needed. The presence of fillers could lead to siphoning, and could improve how to probe your memory, but could also help you set your criterion less variably. Equally likely, the presence of fillers could lead to greater accessibility of diagnostic cues, which in turn could help you set your criterion less variably.

Participants can adjust the way that they treat the memory evidence to suit the task and its demands (McAdoo, Key, & Gronlund, 2018; McAdoo, Key, & Gronlund, 2019). They can either treat the evidence in a continuous manner and carefully weigh the evidence (continualization), or they can treat the evidence in an all-or-none manner and discard the evidence that is not readily apparent (discretization). It would be interesting to determine how participants mediate the memory evidence in these experiments, and whether this precedes setting more consistent criteria across trials. Perhaps, a participant that recognizes the need for more careful deliberation of the evidence (continuous mediation) also requires more evidence overall on which to base an identification decision (sets their criteira more conservatively across trials). In these experiments, participants were generally poor at setting a consistent criteira, and their aforementioned lack of motivation may have prompted them to use a discrete mediation strategy. Unfortunately, we do not have the experimental design to support this kind of exploration; a strength manipulation during encoding is required (Kellen & Klauer, 2015; McAdoo et al., 2019).

Criterial variability was not reduced under these showup testing conditions, but perhaps it is reduced under more realistic testing conditions. In our experiments, a participant's identification accuracy is the average of their many identifications. But in real-world eyewitness scenarios, an eyewitness makes a single identification. Future research should examine the criterial variability hypothesis under conditions that more closely match the eyewitness' task. If our hypotheses are supported in these more scaled-up, realistic scenarios, this would help establish the criterial hypothesis as a valid explanation for the lineup advantage.

References

Amendola, K. L., & Wixted, J. T. (2015). Comparing the diagnostic accuracy of suspect identifications made by actual eyewitnesses from simultaneous and sequential lineups in a randomized field trial. *Journal of Experimental Criminology, 11*(2), 263-284.

Bainbridge, W.A., Isola, P., & Oliva, A. (2013). The intrinsic memorability of face images. *Journal of Experimental Psychology: General*, *142*(4), 1323-1334.

Benjamin, A., Diaz, M., & Wee, S. (2009). Signal detection with criterion noise: applications to recognition memory. *Psychological Review, 116*(1), 84-115.

Brady, T., Robinson, M. M., Williams, J. R., & Wixted, J. (2021). Measuring memory is harder than you think: A crisis of measurement in memory research.

Carlson, C. A., Gronlund, S. D., & Clark, S. E. (2008). Lineup composition, suspect position, and the sequential lineup advantage. *Journal of Experimental Psychology: Applied*, *14*(2), 118–128.

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Routledge.

Cohen, A. L., Starns, J. J., & Rotello, C. M. (2021). sdtlu: An R package for the signal detection analysis of eyewitness lineup data. *Behavior Research Methods, 53*(1), 278-300.

Colloff, M. F., Wade, K. A., Strange, D., & Wixted, J. T. (2019). Filler-siphoning theory does not predict the effect of lineup fairness on the ability to discriminate innocent from guilty suspects: Reply to Smith, Wells, Smalarz, and Lampinen (2018). *Psychological Science*, *29*(9), 1552-1557.

Colloff, M. F., & Wixted, J. T. (2020). Why are lineups better than showups? A test of the filler siphoning and enhanced discriminability accounts. *Journal of Experimental Psychology: Applied*, *26*(1), 124–143.

Florida Department of Corrections Offender Network: http://www.dc.state.fl.us/offenderSearch/

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

Grgic, M., Delac, K., & Grgic, S. (2011). SCface - surveillance cameras face database. *Multimedia Tools and Applications Journal, 51*(3), 863-879.

Gronlund, S. D., & Benjamin, A. S. (2018). Chapter 8- The new science of eyewitness memory. *Psychology of Learning and Motivation, 69,* 241-284.

Gronlund, S. D., Carlson, C. A., Dailey, S. B., & Goodsell, C. A. (2009). Robustness of the sequential lineup advantage. *Journal of Experimental Psychology: Applied, 15*(2), 140–152.

Gronlund, S. D., Carlson, C. A., Neuschatz, J. S., Goodsell, C. A., Wetmore, S. A, Wooten, A., & Graham, M. (2012). Showups versus lineups: An evaluation using ROC analysis. Journal of Applied Research in *Memory and Cognition, 1*(4), 221-228.

Gronlund, S. D., Wixted, J. T., & Mickes, L. (2014). Evaluating eyewitness identification procedures using receiver operating characteristic analysis. *Current Directions in Psychological Science, 23*(1), 3-10.

Innocence Project. (2021). DNA Exonerations in the United States. Retrieved March 19, 2021, from https://innocenceproject.org/dna-exonerations-in-the-united-states/.

Kellen, D., & Klauer, K. C. (2015). Signal detection and threshold modeling of confidence-rating ROCs: A critical test with minimal assumptions. *Psychological Review, 122*(3), 542.

Kellen, D., Klauer, K. C., & Singmann, H. (2012). On the measurement of criterion noise in signal detection theory: The case of recognition memory. *Psychological Review, 119,* 457-479.

Luus, C. A., & Wells, G. L. (1991). Eyewitness identification and the selection of distracters for lineups. *Law and Human Behavior*, *15*(1), 43-57.

Malpass, R. S. (1981). Effective size and defendant bias in eyewitness identification lineups. *Law and Human behavior*, *5*(4), 299.

Malpass, R. S., & Lindsay, R. C. (1999). Measuring lineup fairness. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, *13*(S1), S1-S7.

Malpass, R. S., Tredoux, C. G., & McQuiston-Surrett, D. (2007). Lineup construction and lineup fairness. In *The Handbook of Eyewitness Psychology: Volume II* (pp. 169-192). Psychology Press.

McAdoo, R. M., Key, K. N., & Gronlund, S. D. (2018). Stimulus effects and the mediation of recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(11), 1814.

McAdoo, R. M., Key, K. N., & Gronlund, S. D. (2019). Task effects determine whether recognition memory is mediated discretely or continuously. *Memory & Cognition, 47*(4), 683-695.

McAdoo, R. M. & Gronlund, S. D. Effect of Between-Subject Decision Noise on Eyewitness ROC Analysis: A Theory Space Exploration Using the WITNESS Model. Talk presented at the Psychonomics meetings, Boston, November, 2016.

McQuiston-Surrett, D. E., Malpass, R. S., & Tredoux, C. G. (2006). Sequential vs. simultaneous lineups: A review of method, data, and theory. *Psychology, Public Policy and Law*, *12*, 137–169.

Meissner, C.A., Tredoux, C.G., Parker, J.F., & MacLin O. H. (2005). Eyewitness decisions in

simultaneous and sequential lineups: A dual-process signal detection theory analysis. *Memory & Cognition, 33*(5), 783-792.

Nakagawa, S., Poulin, R., Mengersen, K., Reinhold, K., Engqvist, L., Lagisz, M., & Senior, A. M. (2015). Meta-analysis of variation: ecological and evolutionary applications and beyond. *Methods in Ecology and Evolution*, *6*(2), 143-152.

Psychology Software Tools, Inc. [E-Prime 3.0]. (2016). Retrieved from https://support.pstnet.com/.

Psychology Software Tools, Inc. [E-Prime Go]. (2020). Retrieved from https://support.pstnet.com/.

Ratcliff, R. (2006). Modeling response signal and response time data. *Cognitive Psychology*, *53*(3), 195–237.

Ratcliff, R., & Starns, J. J. (2009). Modeling Confidence and Response Time in Recognition Memory. *Psychological Review*, *116*(1), 59–83.

Smith, A. M., Wells, G. L., Lindsay, R. C. L., & Penrod, S. D. (2017). Fair lineups are better than biased lineups and showups, but not because they increase underlying discriminability. *Law and Human Behavior, 41*(2), 127–145.

Starns, J. J. (2014). Using response time modeling to distinguish memory and decision processes in recognition and source tasks. *Memory & cognition*, *42*(8), 1357-1372.

Steblay, N., Dysart, J., Fulero, S., & Lindsay, R. C. L. (2003). Eyewitness accuracy rates in police showup and lineup presentations: A meta-analytic comparison. *Law and Human Behavior*, *27*(5), 523-540.

Steblay, N. K., Dysart, J. E., & Wells, G. L. (2011). Seventy-two tests of the sequential lineup

superiority effect: A meta-analysis and policy discussion. *Psychology, Public Policy, and Law*, *17*(1), 99.

Tredoux, C. G. (1998). Statistical Inference on Measures of Lineup Fairness. *Law and Human \ Behavior*, *22*(2), 217-237.

Tredoux, C. G. (2018). Statistical Inference on Lineup Fairness: Package 'r4lineups' for R>= \ 3.4.0. Retrieved from: https://cran.microsoft.com/snapshot/2020-04-20/web/packages/r4lineups/r4lineups.pdf

Wells, G. L., Memon, A., & Penrod, S. D. (2006). Eyewitness evidence: Improving its probative value. *Psychological science in the public interest*, *7*(2), 45-75.

Wells, G. L., Smalarz, L., & Andrew, S. M. (2015). ROC analysis of lineups does not measure underlying discriminability and has limited value. *Journal of Applied Research in Memory and Cognition*, *4*(4), 313-317.

Wetmore, S. A., Gronlund, S. D., Neuschatz, J. S., & McAdoo, R. M. Lineups are Better than Showups but Filler Siphoning is Rarely the Reason. Poster presented at the Psychonomics meetings, Boston, November, 2016.

Wetmore, S.A., McAdoo, R.M., Gronlund, S.D., & Neuschatz, J. S. (2017). The impact of fillers on lineup performance. *Cognitive Research: Principles and Implications, 2*(48).

Wixted, J. T., & Mickes L. (2014). A Signal-Detection-Based Diagnostic-Feature-Detection Model of Eyewitness Identification. *Psychological Review, 121*(2), 262–276.

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*(2), 304-309.

Table 1

Experiment 1 Summary Statistics

	Mean d'	*SD d'*	*Mean c*	*SD c*
LU	0.77	0.55	0.77	0.25
SU	0.74	1.11	0.02	0.48

Note. Summary statistics for the standard lineup (LU) and showup (SU) conditions. Mean *d*' denotes the average performance, or the average ability to distinguish between guilty and innocent suspects among participants; *d*' = z(HR) - z(FAR), where z() is the inverse of the cumulative distribution function of the underlying evidence distributions. Mean *c* is the average criterion position adopted by participants; c = –½ [z(HR) + z(FAR)]. Smaller values denote more liberal responding, whereas larger values denote more conservative responding. *Sd*(*c*) is the standard deviation of the group's average criterion placement across trials, and approximates criterial variability.

Table 2

Experiment 2 Summary Statistics (N = 72)

	Mean d'	*SD d'*	*Mean c*	*SD c*
LU (*n* = 24)	0.624	0.416	0.940	0.291
SU (*n* = 24)	0.505	0.358	0.301	0.349
SU_c (*n* = 24)	0.718	0.415	0.703	0.586

Note. Summary statistics for the lineup (LU), showup (SU), and constrained showup (SU_c) conditions. Mean *d*' denotes the average performance among participants. Mean *c* is the average criterion position adopted by participants. *Sd*(*c*) is the standard deviation of the group's average criterion placement across trials and approximates criterial variability.

Table 3

Tredoux's Effective Size (N = 36)

Lineup	Effective Size	Bootstrapped Estimates		
		Bias	SE	CI
1	1.69	0.71	0.74	(-0.50, 2.44)
2	2.24	0.14	0.44	(1.22, 2.92)
3	2.93	0.18	0.50	(1.76, 3.72)
4	2.04	0.68	0.77	(-0.15, 2.89)
5	1.71	0.99	1.02	(-1.21, 2.72)
6	1.51	1.21	1.21	(-2.10, 2.74)
7	3.24	0.16	0.68	(1.76, 4.45)
8	1.12	0.78	0.81	N/A
9	1.49	0.71	0.76	(-0.71, 2.23)
10	3.10	0.23	0.66	(1.59, 4.15)
Average	**2.11**	**0.58**		

Note. Bias is the mean of the bootstrap estimates minus the original statistic. Bootstrapped estimates are based on 1,000 bootstrap replicates. Confidence intervals are estimated at the 95% confidence level.

Table 4

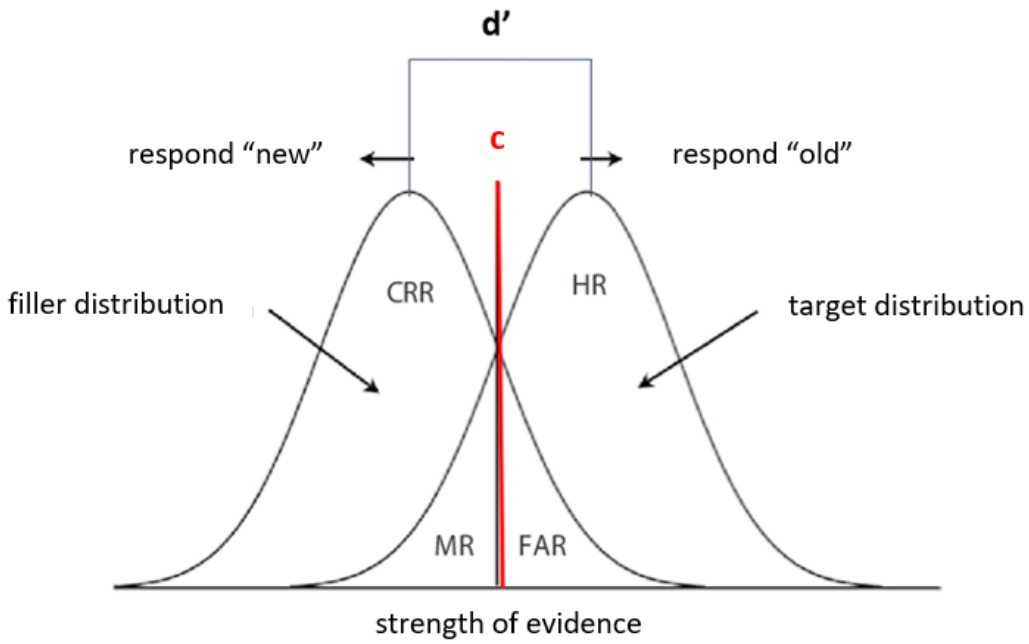
Experiment 3 Summary Statistics (N = 53)

	Mean d'	*SD d'*	*Mean c*	*SD c*
LU (*n* = 24)	1.001	0.599	0.865	0.376
SU (*n* = 17)	0.996	0.557	0.135	0.450
SU_{sim} (*n* = 12)	0.866	0.272	0.284	0.338

Note. Summary statistics for the lineup (LU), showup (SU), and simultaneous showup (SU_{sim}) conditions. Mean *d*' denotes the average performance among participants. Mean *c* is the average criterion position adopted by participants. *Sd*(*c*) is the standard deviation of the group's average criterion placement across trials and approximates criterial variability.

Figure 1

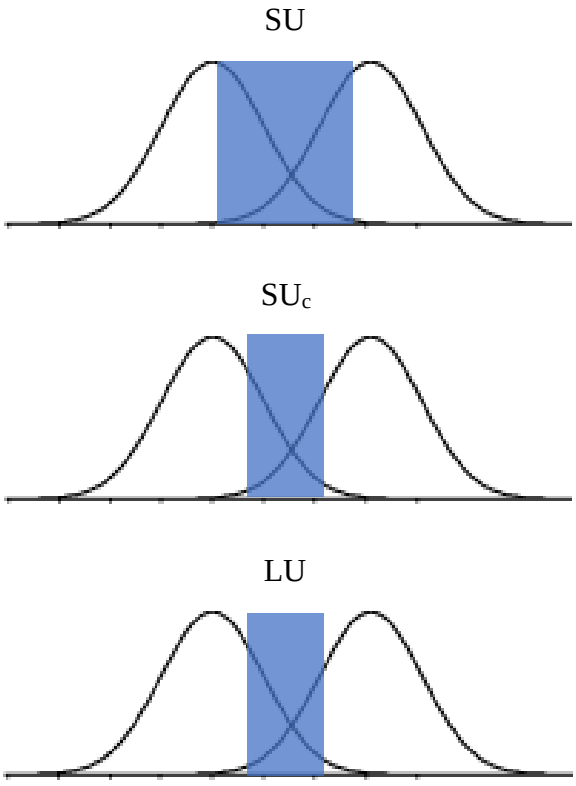
Overlapping Evidence Distributions in Signal Detection Theory



Note. The filler distribution is positioned to the left of the target distribution along the strength of evidence axis. Discriminability is denoted by the d prime (*d'*) measure and is the difference between the mean of the target distribution and the mean of the filler distribution. The response criterion is denoted by *c* and is positioned in the middle of both distributions indicating no differential bias.

Figure 2

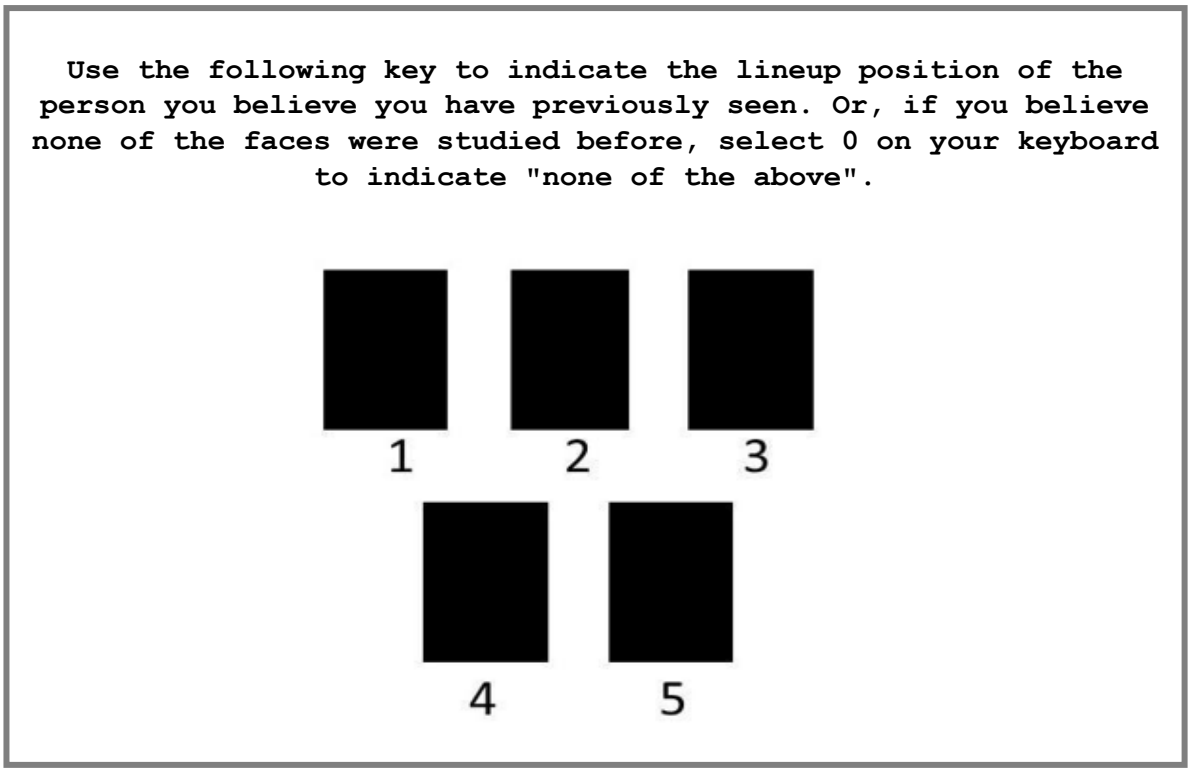
Criterion Placements Across Trials



Note. The shaded blue regions represent the possible ranges of criterion placement across trials for the same individual. Larger regions denote a wider range of possible criterion placements. This may represent an individual who is ill-calibrated to the task or who does not apply consistent criteria for probing their memory evidence, in this case, due to the type of identification procedure. We make these predictions assuming a constant *d'* for the individual.

Figure 3

Spatial Positions (1-5) Representing Faces in a 5-Person Lineup



Note. Black boxes replaced faces and were presented as a response "key" to participants throughout the experiment and beginning in the practice phase, along with the description in bold.

Figure 4

Simultaneous Showup

Note. The suspect is outlined in red. Participants can examine, but not respond, to the non-outlined faces.

Appendix A

Evaluating Lineup Fairness Code

```
#Set working directory
setwd("C:/Users/rebec/Desktop")

# load necessary packages
install.packages("dplyr")
install.packages("readxl")
install.packages("car")
library(dplyr)
library(car)
library(readxl)

# upload data
LU<- as.data.frame(read_excel("dataFA21.xlsx", sheet = "LU"))
SU<- as.data.frame(read_excel("dataFA21.xlsx",sheet = "SU"))

#Analyze the distribution of responses

stim< c()
correct<- c()
for(r in 1:length(TPlus)){ #TPlus stores each individual TP stimulus
  stim[r]<- TPlus[r]
  correct[r]<- as.numeric(LU$correctanswer[which(LU$stim== TPlus[r])][1])
  print(as.numeric(LU$present[which(LU$stim == TPlus[r])]))}

#Gets the total number of correct responses (correct) across subjects for each TP LU
#Exclude on the basis of lowest overall correct responses

response<- c()
for(i in 1:50){
response[i] <- length(which(as.numeric(LU$present[which(LU$stim == TPlus[i])]) == correct[i]))}

stim_TA<- c()
correct_TA<- c()
for(r in 1:length(TAlus)){  #TAlus stores each individual TA stimulus
  stim_TA[r]<- TAlus[r]
  print(as.numeric(LU$present[which(LU$stim == TAlus[r])]))}

#Gets the total number of correct responses (0) across subjects for each TA LU
#Exclude on the basis of lowest overall correct responses

response_TA<- c()
for(i in 1:50){
  response_TA[i] <- length(which(as.numeric(LU$present[which(LU$stim == TAlus[i])]) == 0))}

stim_TASU<- c()
correct_TASU<- c()
for(r in 1:length(TASU)){ #TASU stores each individual TA stimulus
  stim_TASU[r]<- TASU[r]
  print(as.numeric(SU$present[which(SU$stim == TASU[r])]))}

#Gets the total number of correct responses (2) across subjects for each TA LU
#Exclude on the basis of nsub that responded correctly

response_TASU<- c()
for(i in 1:50){
  response_TASU[i] <- length(which(as.numeric(SU$present[which(SU$stim == TASU[i])]) == 2))}
```

Appendix B

Computing Tredoux's E' Code

```
#Set working directory
setwd("C:/Users/rebec/Desktop")

#install.packages("r4lineups")
library(r4lineups)
install.packages("boot")
library(boot)

#Load TP data
lineup_vec<- read.csv("lineup_task.csv")

#Convert to list form for subsequent functions
lineup_vec <- list(lineup_vec[[1]], lineup_vec[[2]],lineup_vec[[3]],lineup_vec[[4]]
                   ,lineup_vec[[5]],lineup_vec[[6]],lineup_vec[[7]],lineup_vec[[8]],
                   lineup_vec[[9]],lineup_vec[[10]])

#Assign nominal size
k<- rep(5, times = 10)

#################### Tredoux's E ##########################################

#Obtain frequency table for each LU
#Compute the effective size on frequency tables

e1 <- esize_T(table(lineup_vec[[1]]))
e2 <- esize_T(table(lineup_vec[[2]]))
e3 <- esize_T(table(lineup_vec[[3]]))
e4 <- esize_T(table(lineup_vec[[4]]))
e5 <- esize_T(table(lineup_vec[[5]]))
e6 <- esize_T(table(lineup_vec[[6]]))
e7 <- esize_T(table(lineup_vec[[7]]))
e8 <- esize_T(table(lineup_vec[[8]]))
e9 <- esize_T(table(lineup_vec[[9]]))
e10 <- esize_T(table(lineup_vec[[10]]))


# compute bootstrapped estimates of effective size
eboot1 <- boot::boot(lineup1_table, esize_T_boot, R = 1000)
eboot2 <- boot::boot(lineup2_table, esize_T_boot, R = 1000)
eboot3 <- boot::boot(lineup3_table, esize_T_boot, R = 1000)
eboot4 <- boot::boot(lineup4_table, esize_T_boot, R = 1000)
eboot5 <- boot::boot(lineup5_table, esize_T_boot, R = 1000)
eboot6 <- boot::boot(lineup6_table, esize_T_boot, R = 1000)
eboot7 <- boot::boot(lineup7_table, esize_T_boot, R = 1000)
eboot8 <- boot::boot(lineup8_table, esize_T_boot, R = 1000)
eboot9 <- boot::boot(lineup9_table, esize_T_boot, R = 1000)
eboot10 <- boot::boot(lineup10_table, esize_T_boot, R = 1000)

#Get confidence intervals using prior bootstrap statistics (bias, se)
ci1 <- boot::boot.ci(eboot1)
ci2 <- boot::boot.ci(eboot2)
ci3 <- boot::boot.ci(eboot3)
ci4 <- boot::boot.ci(eboot4)
ci5 <- boot::boot.ci(eboot5)
ci6 <- boot::boot.ci(eboot6)
ci7 <- boot::boot.ci(eboot7)
ci8 <- boot::boot.ci(eboot8)
ci9 <- boot::boot.ci(eboot9)
ci10 <- boot::boot.ci(eboot10)

########################### LU proportion ################################

#Compute bias in suspect choosing rate
```

```
bias4<- allfoilbias(lineup4_table, 1,5)
bias5<- allfoilbias(lineup5_table, 2,5)
bias6<- allfoilbias(lineup6_table, 4,5)
bias7<- allfoilbias(lineup7_table, 2,5)
bias10<- allfoilbias(lineup10_table, 1,5)

#Compute bootstrapped estimates
#Requires the position of targets
target_pos<- c(5,4,1,1,2,4,2,4,1,1)

for(i in 1:10){
  boot<- boot::boot(lineup_vec[[i]], lineup_prop_boot, target_pos = target_pos[i], R = 1000)# se
  print(boot)
  ci <- boot::boot.ci(boot, conf = 0.95, type = "bca")   #bca = biased corrected CI
  print(ci)
}
```

Appendix C

Supplemental Results: Experiment 2

Tests for Normality

Test	p-value (*d'*)	p-value (*c*)
Shapiro-Wilk	0.521	0.634
Anderson-Darling	0.612	0.641
Cramer-von Mises Test	0.647	0.637

Tests for Homogeneity of Variances

d'		*c*	
Test	p-value	Test	p-value
Bartlett's Test	0.725	Bartlett's Test	< .05
Levene's Test	0.762	Levene's Test	< .05

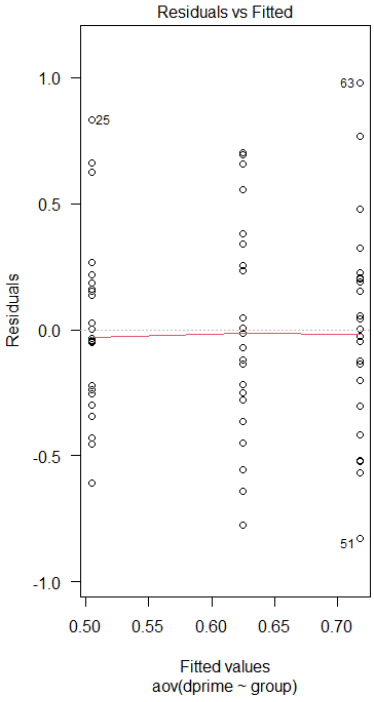
Lineup Proportions (N = 36)

Lineup	Lineup proportion (%)	Bootstrapped Bias	Bootstrapped SE	Bootstrapped CI
1	75	0.0015	0.073	(0.5556, 0.8611)
2	56	-0.003	0.080	(0.3611, 0.6944)
3	47	0.0041	0.083	(0.2500, 0.5902)
4	67	0.0038	0.080	(0.4444, 0.7778)
5	75	0.001	0.070	(0.5556, 0.8611)
6	81	-0.0018	0.064	(0.6354, 0.8889)
7	31	-0.0005	0.076	(0.1389, 0.4167)
8	94	0.00086	0.037	(0.8056, 0.9722)
9	81	-0.0018	0.068	(0.6111, 0.8889)
10	44	-0.0014	0.085	(0.2500, 0.5833)

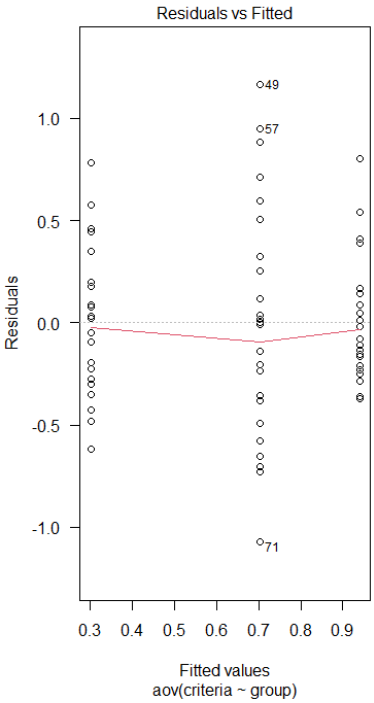
Note. Percentages indicate the choosing rate for any single lineup member for lineups with a designated innocent suspect. Bootstrapped estimates are based on 1,000 bootstrap replicates. Confidence intervals are estimated at the 95% confidence level.

Residual Plots

(a) Discriminability



(b) Criteria



Note. The red line indicates the line of best fit.

Appendix D

Supplemental Results: Experiment 3

Tests for Normality

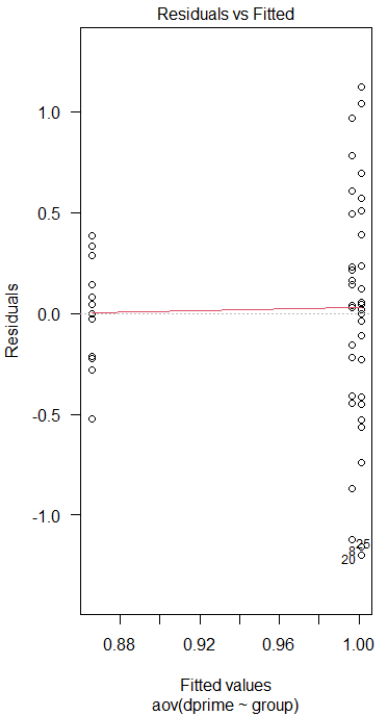
Test	p-value (*d'*)	p-value (*c*)
Shapiro-Wilk	0.633	0.448
Anderson-Darling	0.553	0.685
Cramer-von Mises Test	0.514	0.790

Tests for Homogeneity of Variances

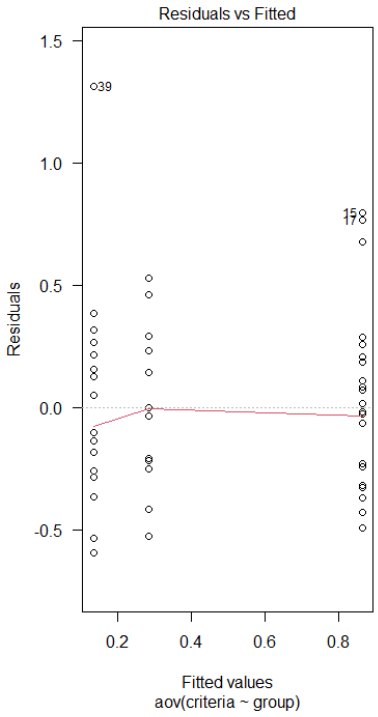
d'		*c*	
Test	p-value	Test	p-value
Bartlett's Test	<.05	Bartlett's Test	0.561
Levene's Test	0.121	Levene's Test	0.841

Residual Plots

(a) Discriminability



(b) Criteria



Note. The red line indicates the line of best fit.

Appendix E

Modal Descriptions (Example)

Directions *In this survey, you will read a description about a face. For each description, please indicate which face in a set of faces you will see best matches the description. You cannot forgo choosing a face.*

Which of the following faces best matches the description below? (Response options 1-5)

1. Caucasian male, late twenties to early thirties, short brown hair, minor facial hair, protruding ears
2. Caucasian male, mid-twenties, medium length brown hair with bangs, thick eyebrows
3. Caucasian female, mid-twenties, medium length off-blond hair, thin eyebrows
4. Caucasian female, mid-twenties, short medium-dark brown hair, thin eyebrows
5. Caucasian male, mid to late twenties, short blonde hairs