ABSOLUTE AND RELATIVE JUDGMENTS AND
THE RELATIONSHIP BETWEEN EYEWITNESS ACCURACY AND CONFIDENCE

A Thesis

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Master of Arts

By

Daniel M. Bialer

August 2019

ABSTRACT

Research in eyewitness identification has found that eyewitness confidence can be highly predictive of eyewitness accuracy if a set of pristine testing conditions are met. Fuzzy-trace theory (FTT), a dual-process theory of cognition and memory, predicts that the distinction between pristine and non-pristine conditions results from differing reliance on verbatim versus gist traces, with verbatim traces used more in pristine conditions and gist traces used more in non-pristine conditions. According to FTT, use of verbatim traces leads to absolute judgments and use of gist traces leads to relative judgments. The current study tests this theory by comparing eyewitness accuracy and confidence for lineups in which the foils are visually similar to the suspect, a requirement for pristine testing, with lineups in which most of the foils are dissimilar to the suspect. We presented these lineups both simultaneously and sequentially. As sequential lineups are thought to promote absolute judgments, we expected that, compared to simultaneous lineups, the distinction in the confidence-accuracy relationship between pristine and non-pristine conditions would be smaller. While we found that adding dissimilar foils to a fair lineup did lead to decreased accuracy and increased confidence in false identifications, we did not find any interactions between the lineup presentation (sequential or simultaneous) and the lineup composition, which does not support our hypothesis. These results may suggest that simultaneous versus sequential lineups may not be an effective manipulation of the use of absolute versus relative judgments. The data suggest that eyewitnesses are making relative judgments from sequential lineups.

Keywords: eyewitness identification, confidence and accuracy, simultaneous and sequential lineups, fuzzy-trace theory

Biographical Sketch

Daniel M. Bialer was born on Long Island on June 5, 1993. He earned his B.A. in Psychology from Vassar College in 2015 and his MSc in Social Cognition from University College London in 2017. From 2015-2016, he worked full-time as a litigation paralegal. Since 2017, he has been working towards a JD/PhD in Law, Psychology, and Human Development at Cornell University.

ACKNOWLEDGEMENTS

I would like to thank the chairperson of my committee, Dr. Charles Brainerd for his help and advice throughout the past two years. His guidance and feedback have not only helped me to write a stronger thesis, but have allowed me to become a better researcher. I would also like to thank my committee members, Dr. Valerie Hans and Dr. Stephen Ceci, for their support in this project. My thesis would also not be possible without the support of the other graduate student and undergraduate student researchers in the Memory and Neuroscience Laboratory. From study design to data collection and analysis, these students have been integral in the completion of this project.

TABLE OF CONTENTS

APPENDIX

LIST OF TABLES

LIST OF FIGURES

Figure page

CHAPTER 1

INTRODUCTION

The purpose of eyewitness identification procedures is to provide law enforcement and jurors with evidence that can help them better assess whether a suspect is guilty or innocent of committing a crime. Therefore, it seems natural to consider additional factors that may help predict the accuracy of an identification. One such factor commonly considered in forensic settings is the eyewitness' self-reported confidence. Confidence has been found to be a highly influential factor in juror assessment of eyewitness accuracy. In fact, it has been found to be more influential than a number of other forensically relevant factors (Cutler, Penrod, & Stuve, 1988).

Despite people's inherent trust in eyewitness confidence, however, it may not always be a reliable predictor of eyewitness accuracy. Of the first 40 DNA exonerations of falsely-convicted individuals in the United States, 90% involved eyewitness misidentifications (Wells et al., 1998). Since 1989, 365 falsely-convicted individuals have been exonerated by DNA evidence and 69% of these false convictions involved eyewitness misidentifications (Innocence Project, 2019). In a study of DNA exonerated cases, Garrett (2011) found that in every case involving eyewitness misidentifications, the eyewitness was highly confident in the identification during the trial. These findings together suggest that eyewitness confidence can be an unreliable predictor of accuracy and that trust in eyewitness confidence can lead to unjust convictions. As an inaccurate conviction is considered a worse outcome to the absence of conviction in the American judicial system (Ceci & Friedman, 2000), these findings are incredibly problematic. In fact, in response, multiple jurisdictions have begun to caution jurors that eyewitness confidence may not be a reliable indicator of accuracy (Wixted & Wells, 2017).

Wixted and Wells (2017), however, propose that confidence *can* be highly predictive of accuracy. In many of the cases in which confidence did not predict accuracy, the procedures that were used were not empirically validated. Wixted and Wells (2017) suggest that if eyewitness identification procedures are conducted under a set of "pristine" conditions, eyewitness confidence can predict accuracy. These pristine conditions include that there should only be one suspect per lineup, that the suspect should not stand out in the lineup, that eyewitnesses should be cautioned that the culprit may or may not be present in the lineup, that the procedure should be conducted double-blindly and that the confidence statements should be taken at the time of the identification.

While we know that confidence can predict accuracy in these pristine conditions, it is not yet clear what theoretically distinguishes pristine conditions from non-pristine conditions. One possible explanation for this distinction stems from fuzzy-trace theory (FTT; Brainerd & Reyna, 2005), a dual-process theory of memory and cognition. According to FTT, memories are stored in parallel as gist and verbatim traces, where gist traces are representations of an event's semantic content and verbatim traces are representations of an event's surface details. FTT predicts that confidence and accuracy will be more closely calibrated when the verbatim traces of the culprit are accessed than when only the gist traces of the culprit are accessed. Consequently, according to FTT we would predict that pristine conditions would promote a greater use of verbatim traces in making eyewitness judgments and that this in turn would lead to a stronger calibration between confidence and accuracy.

While FTT provides a compelling explanation for the distinction between pristine and non-pristine conditions, this application of FTT has never been tested empirically. The current study aims to explicitly test whether FTT can predict the distinction between pristine and non-pristine conditions. Particularly, we aim to test whether FTT can explain why confidence better predicts accuracy in one of the pristine conditions postulated by Wixted and Wells (2017): that the suspect should not stand out from the lineup.

The Effect of Lineup Composition on Confidence and Accuracy

One of the pristine eyewitness identification conditions prescribed by Wixted and Wells (2017) is that the suspect should not stand out in the lineup. Intuitively this makes sense. If for example the perpetrator of a crime has blonde hair and only the suspect in the lineup matches this description, it is logical to assume that an eyewitness would be more confident in his or her decision identifying this suspect than if every other member of the lineup also has blonde hair. In fact, an early assessment of the fairness of a lineup prescribed that if a mock juror could identify the suspect in a lineup simply from the eyewitness' description of the offender then the lineup was unfair (Wells, Leippe, & Ostrom, 1979). Empirical studies have found that eyewitnesses are both more likely to identify suspects and more confident in their identifications when using unfair lineups (Lindsay & Wells, 1980; Wells, Rydell, & Seelau, 1993).

While lineups from which someone can identify a suspect from just an eyewitness' description of the offender are certainly unfair, a lineup can be unfair, but not meet this condition. In fact, simply adding dissimilar foils to a fair lineup can inflate confidence (Charman, Wells, & Joy, 2011). In a study conducted by Charman et al. (2011), participants watched a crime video and were asked to identify the offender from either a simultaneous two-person target-absent lineup or a simultaneous six-person target-absent lineup in which four of the six choices were not plausible. Both lineups had the same two plausible choices and yet participants were significantly more likely to identify innocent suspects from the six-person lineup than from the two-person lineup. Participants also had significantly higher confidence in their identifications from the six-person lineup. This finding is referred to as the "dud-alternative effect" and has also been found in the judgment and decision-making domain (Windschitl & Chamber, 2004) and in other episodic memory tasks (Hanczakowski, Zawadzka, & Higham, 2014). Overall, research suggests that unfair lineup compositions lead to an increase in eyewitness confidence in identifications of innocent suspects, thus weakening the informative value of confidence for predicting accuracy.

A Fuzzy-Trace Theory Account of the Confidence-Accuracy Relationship

The confidence-accuracy relationship has been explained previously using a signal-detection model (Mickes, Hwe, Wais, & Wixted, 2011). According to this model, confidence ratings are representations of various decision criteria existing along a memory strength scale. For example, for a 0-100-point scale, there would be 100 confidence criteria. If the memory strength for a particular suspect succeeded the highest criterion, the eyewitness would rate their confidence as 100. If the memory strength, fell below that criterion, but above the second highest criterion, the eyewitness would rate their confidence as 99. This same pattern would occur for the other points on the confidence scale (Mickes et al., 2011). According to this model, the categorical decision of identifying an individual from a lineup is also based on the same memory strength scale, where an eyewitness would identify an individual if memory strength for that individual exceeds a certain threshold set by the individual. Therefore, we would expect that confidence and accuracy would be closely calibrated according to a signal-detection model (Wixted & Wells, 2017).

The signal-detection model accurately predicts the confidence-accuracy relationship in pristine conditions; however, it does not predict the relationship in non-pristine conditions. The limitation of using a signal-detection model for the confidence-accuracy relationship is that it does not account for the fact that confidence statements may be made based on factors besides the strength of the memory (Wixted & Wells, 2017).

FTT postulates that the reason we see a strong correlation between confidence and accuracy in pristine conditions, but not in non-pristine conditions is that confidence ratings and identification decisions promote different degrees of reliance on verbatim versus gist traces. According to the task calibration principle of FTT, people match the mental representations they use to the demands of a task (Corbin, Reyna, Weldon, & Brainerd, 2015). People often have a preference to use the simplest gist representation that they can within a task's constraints. Resultingly, when people make categorical judgments, they tend to rely more on gist traces than when they make judgments which require an exact numerical response (Corbin et al., 2015).

In episodic memory tasks, people often favor verbatim traces over gist traces because they provide more vivid details (Brainerd & Reyna, 2005). When these verbatim traces are available, we would expect that both tasks which require a categorical response (e.g., identifications) and tasks which require an exact numerical response (e.g., confidence ratings) would be based on the verbatim traces of the eyewitness' memory for the offender of the crime. When verbatim traces are not available, however, people may shift to reliance on gist traces. In this case, the task calibration principle would predict that individuals would rely less on gist for tasks which require an exact numerical response than for tasks which require a categorical response (Brainerd, Nakamura, Reyna, & Holliday, 2017).

Therefore, according to the task calibration principle of FTT, we would predict that confidence ratings would decrease reliance on gist traces for episodic memory tasks. Empirical studies have found evidence to support this prediction of FTT. Firstly, according to this theory, confidence in recognition tasks should be higher for hits than for false alarms. This is because previously studied items contain both verbatim and gist traces and new, but similar items contain only gist traces (Brainerd et al., 2017). A pattern of higher confidence for hits than for false alarms is pervasive in the recognition memory literature (Brainerd & Reyna, 2005). We would also expect that confidence would better reflect accuracy for true memories of previously presented items, which are more likely to provoke verbatim traces, than for correct rejections of new, but similar items. In support of this, confidence has been found to positively predict accuracy for true memories, but negatively predict accuracy for correct rejections (Brainerd et al., 2017; DeSoto & Roediger, 2014; Roediger & DeSoto, 2014). Additionally, according to this theory, people should be able to recall items correctly recognized with high confidence in more vivid detail than items correctly recognized with low confidence. This is because a high confidence identification should represent greater verbatim memory. This has also been found to be the case (Selmeczy & Dobbins, 2014). FTT predicts that confidence will be a better predictor of accuracy when there is access to verbatim traces. Therefore, it predicts that the distinction between pristine and non-pristine conditions is a distinction between whether a condition promotes use of verbatim or gist traces in making identification decisions.

How Fuzzy-Trace Theory Maps onto the Absolute versus Relative Judgment Distinction

Wells (1984) proposed that the distinction between the diagnosticity of eyewitness identifications in pristine and non-pristine conditions could be explained by eyewitnesses' use of relative versus absolute judgments. In pristine conditions, eyewitnesses should be more likely to use an absolute judgment strategy, identifying a lineup member only if memory for that lineup member passes a certain threshold. In contrast, in non-pristine conditions, eyewitnesses may be more likely to use relative judgment strategies, making comparisons between the lineup members and choosing the lineup member who most closely resembles their memory of the offender relative to the other lineup members. From this distinction, it is apparent that more innocent suspects should be incorrectly identified from lineups in non-pristine conditions because they encourage choosing the lineup member who most closely resembles the offender rather than choosing a lineup member only if they match the eyewitness' precise memory for the offender (Wells, 1984).

This distinction between absolute and relative judgment strategies can be mapped onto the distinction between verbatim and gist traces in FTT. An absolute judgment strategy is a verbatim-based comparison, where an eyewitness identifies a member of the lineup only if they match the verbatim details of the offender. In contrast, a relative judgment strategy is a gist-based comparison, where an eyewitness selects the lineup member who most closely resembles his or her memory of the basic features of the offender, such as race, gender, age, and build. In illustrating this difference, it is useful to think of how fair and unfair lineups may lead to the two judgment strategies. In a fair lineup, all the lineup members match the eyewitness' description of the offender. Therefore, in an identification procedure an eyewitness cannot rely on the basic features of the offender to discriminate between the lineup members. In this case, an eyewitness would be more likely to engage in absolute judgments by comparing each lineup member to his or her memory for the offender and only selecting a lineup member who matches the verbatim traces of the memory of the criminal. In contrast, in an unfair lineup where only the suspect matches the eyewitness' description of the offender, eyewitnesses may be more likely to use relative judgments, choosing the suspect based on the fact that they most closely resemble the gist of the offender.

The absolute versus relative judgment explanation of the difference between pristine and non-pristine conditions can also relate to the confidence-accuracy relationship. In a study by

Zawadzka, Higham, and Hanczakowsi (2017), it was found that in a two-alternative forced-choice recognition test, confidence was greater when evidence in support of both the chosen alternative and the non-chosen alternative was stronger. This implies that confidence statements did not reflect the difference in evidence between the two alternatives, but rather that participants were rating their confidence based on the absolute strength of evidence in support of the chosen alternative. This is consistent with the finding that confidence statements are based on verbatim memory in FTT (Brainerd et al., 2017) and suggests that because confidence ratings are based on absolute judgments, they may only reflect categorical decisions which are made based on absolute judgments as well (e.g., when eyewitness identification procedures are conducted under pristine conditions).

Individual Difference Measures and Eyewitness Identification

The eyewitness literature often makes a distinction between system variables, variables under the direct control of the criminal justice system, and estimator variables, which cannot be controlled. Until this section, we have only considered the role of system variables in affecting eyewitness identification; however, it is also important to understand the role of estimator variables. While they are not under the direct control of the criminal justice system, understanding their effects can allow us to better assess the accuracy of identifications. A variety of individual difference measures have been found to affect eyewitness identification performance. These include individual differences in facial recognition (Andersen, Carlson, Carlson, & Gronlund, 2014; Morgan et al., 2007), working memory capacity (Andersen et al., 2014), processing styles (Darling, Martin, Hellmann, & Memon, 2009), age (Searcy, Bartlett, & Memon, 1999) and autistic traits (Andersen et al., 2014).

Overview of the Current Study

The central aim of the current study is to better understand why confidence is more predictive of accuracy in pristine testing conditions than in non-pristine conditions. FTT provides a plausible explanation for this distinction, but it has never been tested empirically. In the current study, we presented participants with crime videos and had them complete identification procedures using fair and unfair lineups, which differed by the extent to which the suspect stood out among the foils. We predicted that participants would be more accurate for fair lineups than for unfair lineups and that participants would be more confident for incorrect identifications from unfair lineups than from fair lineups. We also predicted that confidence and accuracy would be better calibrated for fair than for unfair lineups.

To test the role of FTT in the distinction between fair and unfair lineups we presented the lineups both simultaneously and sequentially. Sequential lineups were designed to limit eyewitness' use of relative judgments in favor of absolute judgments (Lindsay & Wells, 1985). While simultaneous lineups are believed to promote relative judgment strategies (Lindsay & Wells, 1985), people may also make absolute judgments when presented with simultaneous stimuli (Starns, Chen, & Staub, 2017). Therefore, we predicted that in sequential lineups, where participants are more restricted to absolute judgments, we would see a smaller difference in the confidence-accuracy relationship between fair and unfair lineups, than in simultaneous lineups, where participants have more freedom to use both absolute and relative judgments. We also measured individual differences in working memory capacity and suggestibility and predicted that they would affect eyewitness performance.

CHAPTER 2

METHOD

Participants

The participants were 256 undergraduate students at Cornell University who participated in exchange for course credit (153 female, 102 male, 1 unreported gender, M age = 19.5 years, SD = 1.27). 118 participants identified as white, 67 identified as Asian, 29 identified as black, 27 identified as Hispanic or Latino, and the remaining 15 identified with other ethnic backgrounds. One participant, whose accuracy was below chance, was excluded from analysis. An additional five participants had missing operation span data and four participants had missing suggestibility data due to computer and human errors. For all ANCOVA analyses we used mean imputation to accommodate these missing values.

Design

All of the participants were presented with 16 crime videos portraying Caucasian actors followed by eight target present and eight target absent lineups. Participants were randomly assigned to one of three lineup composition conditions (six-person fair lineup, two-person fair lineup, or six-person unfair lineup) and to one of two lineup presentation style conditions (sequential or simultaneous). Therefore, the experiment followed a 2 x 3 x 2 mixed design with a within-subject factor (target presence: target present, target absent) and two between-subject factors (fairness: six-person fair lineup, two-person fair lineup, six-person unfair lineup; and presentation style: sequential or simultaneous).

Materials

Videos. Thirty-two videos of Caucasian young adult actors (16 male, 16 female), created by Mansour et al. (2012), were used as stimuli (approximate size: 560 x 315 pixels). Each video was approximately 30 seconds in length and portrayed an undisguised actor from the shoulders up, facing the camera, and in front of a green background. In half of the videos the actor plans a burglary with an off-camera accomplice and in the other half the actor commits a bank robbery. Half of the participants viewed the videos of the 16 male actors, while the other half viewed the videos of the 16 female actors.

Lineups. The lineups all displayed a selection of photographs of Caucasian faces. Each face portrayed a neutral expression in front of a white background (approximate size:172 x 214 pixels). The faces were undisguised and were displayed from the neck up, so as not to provide any clothing cues. In the simultaneous lineups, the photographs of the lineup members were displayed on a gray background. Each of the lineup members was associated with a number which was displayed above or below the photograph.

The six-person fair lineups were created for Mansour et al. (2012). Target present and target absent lineups were created for each target. Foils were selected from a pool of photographs using an iterative matching process. Mansour et al. (2012) ensured the fairness of the lineups by having a set of participants generate descriptions of the targets and another set of participants try to identify the targets based on their descriptions. Each foil appeared in the lineups for only one of the targets. In the target present lineups, the position of the target was counterbalanced across targets, so that the target would be presented in each position at approximately equal rates. In the target absent lineups, an additional foil was added in place of the target.

The six-person unfair lineups and two-person fair lineups were adapted from the six-person fair lineups. For each unfair lineup, four of the foils were replaced with opposite sex faces: implausible choices for any of the criminals they had witnessed from the videos. This meant that unfair lineups included only two plausible choices: the target and one visually similar foil in target present lineups, and two plausible foils in the target absent lineups. As the six-person fair lineups had six plausible choices, but the unfair lineups had only two, two-person fair lineups were also used as a second control. The two-person fair lineups were identical to the unfair lineups except that they did not include the four opposite sex foils. This allowed for a direct comparison of the effect of adding dissimilar foils to a lineup. The mean position of the culprit in the target present six-person fair lineups and target present six-person unfair lineups was 3.53 (3.56 for the male stimuli and 3.50 for the female stimuli). The mean position of the culprit in the target present two-person fair lineups was 1.50 for both the male and female stimuli. Appendix A displays an example of a six-person fair, six-person unfair, and two-person fair lineup for the same target.

Lineup Instructions. Participants read and listened to instructions before beginning the lineup portion of the study. They were told that they would be presented with a series of lineups from which they would be asked to identify the criminals from the videos that they had watched earlier. They were told that the order of the lineups would not correspond to the order in which they had seen the videos and that the lineups may or may not include one of the criminals from the videos.

For simultaneous lineups, participants were told to either identify the number corresponding to the photograph of the person who they believed they had observed from one of the videos or to select “none of the above” for six-person lineups and “neither of the above” for two-person lineups if they believed that none of the criminals observed from the videos were present in the lineup.

For sequential lineups, participants were told to either choose “yes” if they believed that they had observed the currently presented lineup member from one of the videos or “no” if they believed they had not observed the lineup member. They were told that if they chose “yes” the current lineup would end and that if they chose “no” they would advance to the next member of the lineup. It was explained that they would continue advancing through a lineup until they had either selected “yes” for a lineup member or selected “no” for every member of the lineup.

Participants in both the simultaneous and sequential conditions were then told that after each lineup, they would rate their confidence in their decision on a 0 to 100 scale, where 100 represented complete confidence and 0 represented a random guess.

Operation Span Task. An operation span task coded in Python by von der Malsburg (2015) was administered to all participants as a measure of working memory capacity. During the task, participants were presented with a series of mathematical equations with two arithmetic operations on the left side of the equation and a stated solution on the right side. For each equation, participants had to determine whether the equation was correct or incorrect. Between the math equations, consonant letters would appear on the screen. After two, three, four, or five letters were presented, participants were tested on their ability to remember the letters. Participants received a total of 12 tests, three each for the sets of two, three, four and five letters. Working memory capacity was calculated based on the number of letters the participants had correctly recalled.

Gudjonsson Suggestibility Scale. The Gudjonsson Suggestibility Scale (GSS 1), a measure of individual differences in suggestibility (Gudjonsson, 1984a) was administered to all participants. During the GSS 1, participants listened to an audio recording of a crime story, completed written free recall, and were asked a series of 20 questions by the experimenter. 15 of the 20 questions were suggestive in nature. Participants were then given negative feedback about their performance on the questions and asked to answer the 20 questions a second time. Free recall accuracy was measured by counting the number of correctly recalled elements from the story out of 40. Suggestion was measured by taking the sum of the number of affirmative answers to suggestive questions during the first round of questioning, known as "yield", and the number of answers that the participant changed between the first and second rounds of questioning, known as "shift". The GSS 1 has been found to be a valid measure of suggestibility in both forensic samples (Gudjonsson, 1984b) and samples of undergraduates (Merckelbach, Muris, Wessel, & Van Koppen, 1998). As the GSS 1 was written in the United Kingdom, we adapted the original scale for an American audience (see Appendix B for comparison).

Procedure

Participants were randomly assigned to either receive simultaneous or sequential lineups. Within each lineup presentation style participants were randomly assigned to one of the three lineup composition conditions: six-person fair, six-person unfair, or two-person fair lineups. Participants were also randomly assigned to either watch videos of female criminals or male criminals. Within each of these 12 combinations, participants were randomly assigned to one of four sub-conditions, which differed based on which of the lineups were target present versus target absent.

Participants were tested individually using Qualtrics survey software and Python programming language. They were first told that they would be presented with a series of videos that they should pay close attention to as they would later be tested on the identity of the individuals from the videos. Participants then watched the 16 crime videos (either all male or all female). After participants watched all the videos, they completed the operation span task. This acted as both a buffer task and as a measure of individual differences in working memory capacity. Following the operation span task, participants read and listened to the lineup instructions and completed 16 lineups (eight target present and eight target absent), which each corresponded to one of the crime videos that they had watched. The lineups were presented in a random order.

Participants in the simultaneous condition were presented with all the members of the lineup at once. For each lineup, participants either chose one of the numbers corresponding to a lineup member if they believed that they had observed that lineup member from one of the crime videos or “none of the above” for six-person lineups and “neither of the above” for two-person lineups if they believed that none of the members of the lineup had been present in any of the videos.

Participants in the sequential condition were presented with the lineup members one-by-one. For each lineup member, participants chose either “yes” if they believed they had witnessed that lineup member in one of the crime videos or “no” if they believed that they had not witnessed that lineup member. If a participant chose “yes” that particular lineup would end and if they chose “no” they would advance to the next member of the lineup. If they chose “no” for every member of the lineup, the lineup would also end, as this would be comparable to a selection of “none of the above” on a simultaneous lineup.

After completing each lineup, participants were presented with a slide bar on a scale from 0-100 and asked to rate their degree of confidence in their decision. Once they reached the confidence rating portion of a particular lineup, they could not return to change their decision.

Finally, after completing all 16 lineups, participants completed the GSS 1, as a measure of individual differences in suggestibility.

CHAPTER 3

RESULTS

The ANCOVAs that we used in our analyses included two individual difference measures: working memory capacity and suggestibility. Working memory capacity was measured by taking the average proportion of letters correctly recalled for the 12 test trials in the operation span task. The mean working memory capacity value was 0.82 (SD = 0.13). Suggestibility was measured by adding together the number of suggestive questions that a participant yielded to (M = 3.41, SD = 2.60) and the number of answers participants shifted after receiving negative feedback of their performance (M = 3.81, SD = 3.09) in the GSS 1 interview component. The yield scores could range from 0 to 15 and the shift scores could range from 0 to 20. We also measured the number of story elements (out of 40) that participants recalled from the GSS 1 story (M = 21.2, SD = 6.69), but these scores were not used in our analyses.

There are three possible responses to target present lineups (correct identification, foil identification, and incorrect rejection of the lineup) and two possible responses to target absent lineups (correct rejection and false alarm). Table 1 displays the rates at which participants made each of these responses. The results are then divided into sections which each focus on one of four dependent variables: accuracy, choosing rate, confidence, and the relationship between accuracy and confidence.

Table 1

Identification Performance by Lineup Presentation and Lineup Fairness Conditions

		Target Present Lineups						Target Absent Lineups			
		Correct Identifications		Foil Identifications		Incorrect Rejections		Correct Rejections		False Alarms	
Presentation	**Fairness**	**M**	**SD**	**M**	**SD**	**M**	**SD**	**M**	**SD**	**M**	**SD**
Simultaneous	Six-Person Fair	0.60	0.24	0.25	0.19	0.15	0.16	0.44	0.24	0.56	0.24
Simultaneous	Six-Person Unfair	0.73	0.18	0.11	0.17	0.16	0.14	0.56	0.24	0.44	0.24
Simultaneous	Two-Person Fair	0.77	0.19	0.08	0.11	0.15	0.16	0.65	0.22	0.35	0.22
Sequential	Six-Person Fair	0.47	0.19	0.38	0.24	0.15	0.17	0.42	0.30	0.58	0.30
Sequential	Six-Person Unfair	0.61	0.20	0.15	0.13	0.24	0.20	0.60	0.29	0.40	0.29
Sequential	Two-Person Fair	0.71	0.17	0.08	0.11	0.21	0.16	0.69	0.23	0.31	0.23

Accuracy Measures

When assessing the accuracy of different identification procedures, it is necessary to consider both the sensitivity and specificity of each procedure. In recognition memory research, it is common to generate memory discrimination indices, which consider both hits (correct acceptances of observed items) and false alarms (incorrect acceptances of unobserved items). Such indices have also been used in eyewitness identification research (e.g., Dobolyi & Dodson, 2013). Two of the most commonly used indices are *d'* and P_r (see Snodgrass & Corwin, 1988). While these measures have been found to be highly correlated (e.g., Seamon et al., 2002), differences between them have also been found. For example, Snodgrass and Corwin (1988) found that P_r was more sensitive to changes in bias and discrimination than *d'*. Therefore, we included both measures in our analyses.

***d'* analysis.** *d'* is the discrimination index used for signal detection theory. According to signal detection theory, recognition memory exists on a single scale of familiarity, where an item is accepted as old if it exceeds a certain level of familiarity (Snodgrass & Corwin, 1988). *d'* is calculated by subtracting the z-score of the false alarm rate in target absent lineups from the z-score of the hit rate in target present lineups (Snodgrass & Corwin, 1988; Stanislaw & Todorov, 1999). In this experiment, we did not designate any of the foils in the target absent lineup as the innocent suspect. As a result, in order to calculate the false alarm rate, we divided the number of foil identifications in the target absent lineup by the number of plausible choices in that lineup (2 for two-person fair lineups and six-person unfair lineups, and 6 for six-person fair lineups). *d'* is undefined when the hit rate or the false alarm rate is zero, so we added 0.1 to the numerators and 0.2 to the denominators as a conservative transformation of the hit rates and false alarms. For example, if a participant correctly identified the culprit in 3 out of 8 target present lineups, we would transform this to 3.1/8.2 when calculating the hit rate.

Table 2 displays the d' values by condition. We ran a 3 (fairness: six-person fair, six-person unfair, two-person fair) X 2 (presentation style: simultaneous, sequential) ANCOVA on *d'* scores. The two covariates we used were working memory capacity and suggestibility. All mean and standard deviation estimates were calculated at the mean values of the covariates. The ANCOVA revealed a main effect of lineup fairness, $F(2, 247) = 4.32$, $p = .014$, $\eta_p^2 = 0.034$. Bonferroni-corrected post-hoc tests revealed that *d'* scores were significantly higher for two-person fair lineups (M = 1.90, SD = 0.95) than for six-person unfair lineups (M = 1.53, SD = 0.95), p = .038 and for six-person fair lineups (M = 1.53, SD = 0.95), p = .032. The ANCOVA also revealed a main effect of lineup presentation style, $F(1, 247) = 7.01$, $p = .009$, $\eta_p^2 = 0.028$. *d'* scores were significantly higher for simultaneous (M = 1.81, SD = 0.94) than for sequential (M = 1.50, SD = 0.94) lineups. There was not a significant interaction between lineup fairness and lineup presentation, $F(2, 247) = 0.39$, $p = .68$, $\eta_p^2 = 0.003$. Additionally, there was a significant effect of suggestibility on *d'* scores, $F(1, 247) = 3.95$, $p = .048$, $\eta_p^2 = 0.016$. Higher suggestibility scores were associated with lower *d'* scores.

***P_r* analysis.** P_r is the discrimination index used for two-high threshold theory. According to this theory there are two separate memory thresholds, one for oldness judgments for old items and one for newness judgments for new items. There are therefore two discrimination indices: P_o is the probability that an old item will exceed the oldness threshold and P_n is the probability that a new item will exceed the newness threshold. As these two separate thresholds cannot be determined from single hit and false alarm rates, we make the assumption that the thresholds are equal and compute a single discrimination index: P_r (Snodgrass & Corwin, 1988). P_r is calculated by subtracting the false alarm rate from the hit rate. As we did for *d'*, we calculated the false alarm rate by dividing the number of foil identifications in target absent lineups by the number of plausible choices. If the difference between the hit rate and false alarm rate was negative, we set the P_r rate to 0.

Table 2 displays the P_r values by condition. We ran a 3 (fairness: six-person fair, six-person unfair, two-person fair) X 2 (presentation style: simultaneous, sequential) ANCOVA on P_r scores with working memory capacity and suggestibility as covariates. The ANCOVA revealed a main effect of lineup fairness, *F*(2, 247) =7.60, p = .001, $\eta_p^2 = 0.058$. Bonferroni-corrected post-hoc tests revealed that P_r scores were significantly higher for two-person fair lineups (M = 0.57, SD = 0.23) than for six-person unfair lineups (M = 0.47, SD = 0.23), p = .008 and for six-person fair lineups (M = 0.45, SD = 0.23), p = .001. The ANCOVA also revealed a main effect of lineup presentation style, *F*(1, 247) = 11.1, p = .001, $\eta_p^2 = 0.043$. P_r scores were significantly higher for simultaneous (M = 0.54, SD = 0.23) than for sequential (M = 0.45, SD = 0.24) lineups. There was not a significant interaction between lineup fairness and lineup presentation, *F*(2, 247) = 0.69, p = .50, $\eta_p^2 = 0.006$. These results were consistent with those found using *d'*. The suggestibility score, however, did not significantly affect the P_r score.

Table 2

Accuracy Measures by Lineup Presentation and Lineup Fairness Conditions

		d'		***P_r***	
Presentation	**Fairness**	**M**	**SD**	**M**	**SD**
Simultaneous	Six-Person Fair	1.74	1.04	0.51	0.27
Simultaneous	Six-Person Unfair	1.63	0.94	0.51	0.22
Simultaneous	Two-Person Fair	2.03	1.06	0.60	0.24
Sequential	Six-Person Fair	1.31	0.74	0.38	0.21
Sequential	Six-Person Unfair	1.41	1.03	0.41	0.25
Sequential	Two-Person Fair	1.81	0.86	0.55	0.20

Choosing Rate Measures

While the discrimination indices can tell us how effectively participants can discriminate between the criminals in the videos and innocent foils, they do not tell us how likely participants are to identify someone from a lineup. Participants in all conditions were given the opportunity to choose to not identify any of the lineup members. Therefore, there is naturally variation in the memory thresholds that individuals may set before they make an identification. Signal detection theory and two-high threshold theory each have their own choosing rate measures: *C* and B_r respectively, which are used to estimate the memory criterion.

***C* analysis**. Criterion *C* is a measure of the choosing rate for signal detection theory models. *C* is calculated by adding together the z-score of the hit rate and of the false alarm rate and dividing by -2 (Snodgrass & Corwin, 1988; Stanislaw & Todorov, 1999). We employed the same corrections in this case as we did for *d'*: for the false alarm rate, we divided the number of foil identifications by the number of plausible lineup members and we employed a transformation in which we added 0.1 to the numerators and 0.2 to the denominators for the hit rate and the false alarm rate. Higher scores on this measure represent more conservative choosing of lineup members.

Table 3 displays the *C* scores by condition. We ran a 3 (fairness: six-person fair, six-person unfair, two-person fair) X 2 (presentation style: simultaneous, sequential) ANCOVA on *C* scores with working memory capacity and suggestibility as covariates. The ANCOVA revealed a main effect of lineup fairness, *F*(2, 247) = 29.2, p < .001, $\eta_p^2 = 0.19$. Bonferroni-corrected post-hoc tests revealed that *C* scores were significantly higher for six-person fair lineups (M = 0.61, SD = 0.42) than for six-person unfair lineups (M = 0.22, SD = 0.42), p < .001 and for two-person fair lineups (M = 0.16, SD = 0.42), p < .001, which suggests that participants were more likely to identify a lineup member from six-person unfair lineups and two-person fair lineups than from six-person fair lineups. The ANCOVA also revealed a main effect of lineup presentation style, *F*(1, 247) = 18.2, p < .001, $\eta_p^2 = 0.069$. *C* scores were significantly higher for sequential (M = 0.44, SD = 0.42) than for simultaneous (M = 0.22, SD = 0.42) lineups. There was not a significant interaction between lineup fairness and lineup presentation, *F*(2, 247) = 0.50, p = .61, $\eta_p^2 = 0.004$. There was also a significant effect of suggestibility on *C* scores, *F*(1, 247) = 13.0, p < .001, $\eta_p^2 = 0.050$. Higher suggestibility scores were associated with lower *C* scores, suggesting that more suggestible participants also had higher choosing rates.

B

Table 3 displays the B_r scores by condition. We ran a 3 (fairness: six-person fair, six-person unfair, two-person fair) X 2 (presentation style: simultaneous, sequential) ANCOVA on B_r scores with working memory capacity and suggestibility as covariates. All of the B_r results were consistent with the *C* results. The ANCOVA revealed a main effect of lineup fairness, *F*(2, 247) = 24.5, $p < .001$, $\eta_p^2 = 0.17$. Bonferroni-corrected post-hoc tests revealed that B_r scores were significantly lower for six-person fair lineups (M = 0.21, SD =0.21) than for six-person unfair lineups (M = 0.41, SD = 0.22), $p < .001$ and for two-person fair lineups (M = 0.42, SD = 0.21), $p < .001$, which suggests that participants were more likely to identify a lineup member from six-person unfair lineups and two-person fair lineups than from six-person fair lineups. The ANCOVA also revealed a main effect of lineup presentation style, *F*(1, 247) = 14.0, $p < .001$, $\eta_p^2 = 0.054$. B_r scores were significantly higher for simultaneous (M = 0.40, SD = 0.22) than for sequential (M = 0.30, SD = 0.21) lineups. There was not a significant interaction between lineup fairness and lineup presentation, *F*(2, 247) = 0.55, p = .58, $\eta_p^2 = 0.004$. There was also a significant effect of suggestibility on B_r scores, *F*(1, 247) = 12.5, $p < .001$, $\eta_p^2 = 0.048$. Higher suggestibility scores were associated with higher B_r scores, suggesting that more suggestible participants also had higher choosing rates.

Table 3

Choosing Rate Measures by Lineup Presentation and Lineup Fairness Conditions

		C		*B_r*	
Presentation	**Fairness**	**M**	**SD**	**M**	**SD**
Simultaneous	Six-Person Fair	0.51	0.38	0.25	0.19
Simultaneous	Six-Person Unfair	0.07	0.44	0.48	0.23
Simultaneous	Two-Person Fair	0.06	0.46	0.47	0.28
Sequential	Six-Person Fair	0.72	0.35	0.17	0.13
Sequential	Six-Person Unfair	0.34	0.48	0.35	0.22
Sequential	Two-Person Fair	0.28	0.44	0.36	0.24

Confidence Measures

In every trial, participants assessed their confidence on a 0 to 100 scale, with 100 representing complete confidence in their decision and 0 representing a random guess. It is important to acknowledge, however, that these confidence ratings were taken after five different possible identification decisions. For target present lineups, participants could either correctly identify a culprit from one of the videos (correct acceptance), falsely identify an innocent foil (foil identification), or incorrectly reject the lineup (incorrect rejection). For target absent lineups, participants could either correctly reject the lineup (correct rejection) or falsely identify an innocent foil (false alarm). Resultantly, lineup presentation style and fairness may have different effects on confidence based on the differential identification decisions. Therefore, we ran ANCOVAs separately for each of these five identification decisions. For each identification decision, we ran a 3 (fairness: six-person fair, six-person unfair, two-person fair) X 2 (presentation style: simultaneous, sequential) ANCOVA on confidence scores with working memory capacity and suggestibility as covariates. Confidence ratings for each of the five decision types are displayed by condition in Table 4.

Correct identification confidence. For correct identifications of culprits in target present lineups, there were no significant main effects for lineup fairness, $F(2, 246) = 2.51$, $p = .083$, $\eta_p^2 = 0.020$, nor for lineup presentation style, $F(1, 246) = 0.56$, $p = .46$, $\eta_p^2 = 0.002$. There was also not a significant interaction between fairness and lineup presentation style, $F(2, 246) = 0.32$, $p = .73$, $\eta_p^2 = 0.003$.

Foil identification confidence. For foil identifications in target absent lineups, the ANCOVA revealed a main effect of lineup fairness, $F(2, 146) = 3.49$, $p = 0.033$, $\eta_p^2 = 0.046$. Bonferroni-corrected post-hoc tests revealed that confidence was significantly higher for six-person unfair lineups (M = 57.5, SD = 20.6) than for six-person fair lineups (M = 47.8, SD = 19.6), p = .038. The ANCOVA also revealed a main effect of lineup presentation style, *F*(1, 146) = 10.5, p = .002, $\eta_p^2 = 0.067$. Confidence in foil identifications was significantly higher for sequential lineups (M = 56.6, SD = 20.3) than for simultaneous lineups (M = 45.8, SD = 20.6). There was not a significant interaction between lineup fairness and lineup presentation, *F*(2, 146) = 0.30, p = .74, $\eta_p^2 = 0.004$.

Incorrect rejection confidence. For incorrect rejections of target present lineups, the ANCOVA did not reveal a significant main effect for fairness, *F*(2, 167) = 2.83, p = .062, $\eta_p^2 = 0.033$, nor a significant interaction between fairness and presentation style, *F*(2, 167) = 0.088, p = .92, $\eta_p^2 = 0.001$. There was, however, a significant main effect of presentation style, *F*(1, 167) = 6.94, p = .009, $\eta_p^2 = 0.040$. Confidence in incorrect rejections was significantly higher for sequential lineups (M = 58.3, SD = 22.4) than for simultaneous lineups (M = 49.3, SD = 22.6).

Correct rejection confidence. For correct rejections of target absent lineups, the ANCOVA revealed a significant main effect of lineup fairness, *F*(2, 235) = 11.4, p < .001, $\eta_p^2 = 0.088$. Bonferroni-corrected post-hoc tests revealed that confidence was significantly higher for six-person unfair lineups (M = 65.1, SD = 19.6) than for six-person fair lineups (M = 50.2, SD = 19.6), p < .001, or for two-person fair lineups (M = 56.5, SD = 19.6), p = .015. The ANCOVA also revealed a significant main effect of lineup presentation style, *F*(1, 235) = 12.8, p < .001, $\eta_p^2 = 0.051$. Confidence in correct rejections was significantly higher for sequential lineups (M = 61.7, SD = 19.6) than for simultaneous lineups (M = 52.8, SD = 19.6). There was not a significant interaction between lineup fairness and lineup presentation, *F*(2, 235) = 0.68, p = .51, $\eta_p^2 = 0.006$. There was also a significant effect of suggestibility on confidence in correct rejections, $F(1, 235) = 4.00$, $p = .047$, $\eta_p^2 = 0.017$. Higher suggestibility scores were associated with lower confidence in correct rejections.

False alarm confidence. For foil identifications from target absent lineups, the ANCOVA revealed a significant main effect of lineup fairness, $F(2, 227) = 7.45$, $p = .001$, $\eta_p^2 = 0.062$. Bonferroni-corrected post-hoc tests revealed that confidence was significantly higher for six-person unfair lineups (M = 57.3, SD = 17.3) than for six-person fair lineups (M = 46.7, SD = 17.2), $p < .001$. The ANCOVA also revealed a significant main effect of lineup presentation style, $F(1, 227) = 7.54$, $p = .007$, $\eta_p^2 = 0.032$. Confidence in foil identifications was significantly higher for sequential lineups (M = 54.7, SD = 17.2) than for simultaneous lineups (M = 48.5, SD = 17.2). There was not a significant interaction between lineup fairness and lineup presentation, $F(2, 227) = 0.066$, $p = .94$, $\eta_p^2 = 0.001$.

Table 4

Confidence by Lineup Presentation and Lineup Fairness Conditions

		Target Present Lineups						Target Absent Lineups			
		Correct Identifications		Foil Identifications		Incorrect Rejections		Correct Rejections		False Alarms	
Presentation	**Fairness**	**M**	**SD**	**M**	**SD**	**M**	**SD**	**M**	**SD**	**M**	**SD**
Simultaneous	Six-Person Fair	74.4	18.0	42.7	19.5	45.9	23.0	47.0	23.0	43.5	13.1
Simultaneous	Six-Person Unfair	78.1	13.1	53.1	24.0	54.9	23.9	61.0	18.5	53.9	20.5
Simultaneous	Two-Person Fair	75.7	15.7	41.3	14.5	47.3	21.9	49.7	17.1	48.1	18.0
Sequential	Six-Person Fair	73.8	19.8	52.9	19.5	56.5	20.8	53.8	20.9	50.1	14.9
Sequential	Six-Person Unfair	80.8	13.5	61.6	20.7	64.0	19.9	68.3	19.5	60.6	17.6
Sequential	Two-Person Fair	77.8	13.8	55.4	17.1	54.3	22.8	63.7	18.5	53.2	17.8

Confidence and Accuracy

Our central research question was whether the difference in the confidence-accuracy relationship between pristine and non-pristine questions would be greater for simultaneous lineups than for sequential lineups. Therefore, naturally, we included measures that examined confidence and accuracy together. To do this we used three measures: Somers' D, confidence-accuracy characteristic (CAC) curves, and receiver operating characteristic (ROC) curves.

Somers' D analysis. Somers' D is a measure of ordinal association that can be used to measure the relationship between confidence and accuracy (e.g., Dobolyi & Dodson, 2013). To calculate the Somers' D values for choosers (those who made an identification from a lineup) we first binned the confidence ratings into five categories: 0-29, 30-49, 50-69, 70-89, and 90-100. We then counted the number of correct identifications from target present lineups and the number of false identifications from target absent lineups at each of these five confidence levels for each participant. We generated a concordant pair value by multiplying the number of correct identifications at each confidence level by the number of false alarms at each of the confidence levels below it and adding all of these values together. We then generated a discordant pair value by multiplying the number of correct identifications at each confidence level by the number of false alarms at each of the confidence levels above it and adding all of these values together. We also generated a tied pairs value by multiplying the number of correct identifications with the number of false alarms at each confidence level. Finally, we generated a single number representing the confidence-accuracy relationship by subtracting the discordant pairs value from the concordant pairs value and dividing by the sum of the concordant pairs value, the discordant pairs value, and the tied pairs value (Pannu & Kaszniak, 2005, Somers, 1962). For participants who made no correct identifications or no false alarms, Somers' D is undefined, so in these cases we filled in the mean for these missing values. We also used the same calculations with correct rejections and incorrect rejections in order to generate Somers' D values for non-choosers.

Somers' D scores for choosers and non-choosers are displayed by condition in Table 5. We first ran a 3 (fairness: six-person fair, six-person unfair, two-person fair) X 2 (presentation style: simultaneous, sequential) ANCOVA on the chooser Somers' D scores with working memory capacity and suggestibility as covariates. The ANCOVA did not reveal a main effect of lineup fairness, *F*(2, 247) = 2.29, p = .10, $\eta_p^2 = 0.018$, nor a main effect of lineup presentation style, *F*(1, 247) = 0.44, p = .51, $\eta_p^2 = 0.002$. There was also not a significant interaction between lineup fairness and lineup presentation style, *F*(2, 247) = 0.13, p = .88, $\eta_p^2 = 0.001$. We also ran a 3 (fairness: six-person fair, six-person unfair, two-person fair) X 2 (presentation style: simultaneous, sequential) ANCOVA on the non-chooser Somers' D scores with working memory capacity and suggestibility as covariates. The ANCOVA also did not reveal a main effect of lineup fairness, *F*(2, 247) = 1.34,p = .27, $\eta_p^2 = 0.011$, nor a main effect of lineup presentation style, *F*(1, 247) = 0.23, p = .63, $\eta_p^2 = 0.001$. The ANCOVA also did not reveal a significant interaction, *F*(2, 247) = 1.91, p = .15, $\eta_p^2 = .015$.

Table 5

Somers' D Values by Lineup Presentation and Lineup Fairness Conditions

		Choosers		***Non-Choosers***	
Presentation	**Fairness**	**M**	**SD**	**M**	**SD**
Simultaneous	Six-Person Fair	0.61	0.42	0.08	0.40
Simultaneous	Six-Person Unfair	0.49	0.38	0.08	0.50
Simultaneous	Two-Person Fair	0.60	0.30	0.05	0.42
Sequential	Six-Person Fair	0.55	0.42	0.02	0.39
Sequential	Six-Person Unfair	0.47	0.37	0.03	0.43
Sequential	Two-Person Fair	0.59	0.31	0.25	0.51

Confidence-accuracy characteristic curves. CAC curves are a graphical representation of the confidence-accuracy relationship. They are considered particularly relevant in a forensic setting due to the fact that they generally represent the accuracy of an identification given that an eyewitness identified the suspect in a lineup at each confidence level (Wixted & Wells, 2017). To generate the CAC curves for choosers, we first binned confidence into the same five categories as we did for the Somers' D: 0-29, 30-49, 50-69, 70-89, and 90-100. We then added up the number of correct identifications from target present lineups and the number of false identifications from target absent lineups for each confidence bin aggregated across participants. Next, we divided the number of false identifications from target absent lineups by the number of plausible choices (2 in six-person unfair lineups and two-person fair lineups and 6 in six-person fair lineups) to generate a false alarm rate because we did not designate any innocent suspect from the target absent lineups. The reason this is necessary is because in a correctly designed lineup there should only be one suspect among a set of known innocent foils. Therefore, the comparison that law enforcement is most interested in, is how likely confidence is to predict accuracy given that the suspect was identified. By dividing by the number of plausible lineup members we get an approximation of innocent suspect identifications in the target absent lineup. We next divided the number of correct identifications by the sum of the number of correct identifications and the number of innocent suspect identifications, which provided us a single number for each confidence bin. This number represents the rate of correct identifications given that an eyewitness identified a suspect. As we had equal numbers of target present and target absent lineups, the chance rate for randomly guessing should be 0.5 in this case (Wixted & Wells, 2017).

The CAC curves for choosers across the six conditions can be seen in Figure 1. On the x-axis we included the five confidence bins and on the y-axis we plotted the rate at which participants who identified a suspect were correct in their identifications. The dotted line represents a perfectly calibrated confidence-accuracy relationship where eyewitnesses who identify a suspect at 0-29 confidence are making a random guess and correctly identifying a guilty suspect 50% of the time that they identify a suspect, whereas eyewitnesses who identify a suspect at 90-100 confidence are making an identification with complete certainty and correctly identifying a guilty suspect 100% of the time that they identify a suspect. Points above this dotted line represent accuracy above what would be expected at that confidence level, whereas points below this line represent accuracy below what would be expected at that confidence level.

Figure 1

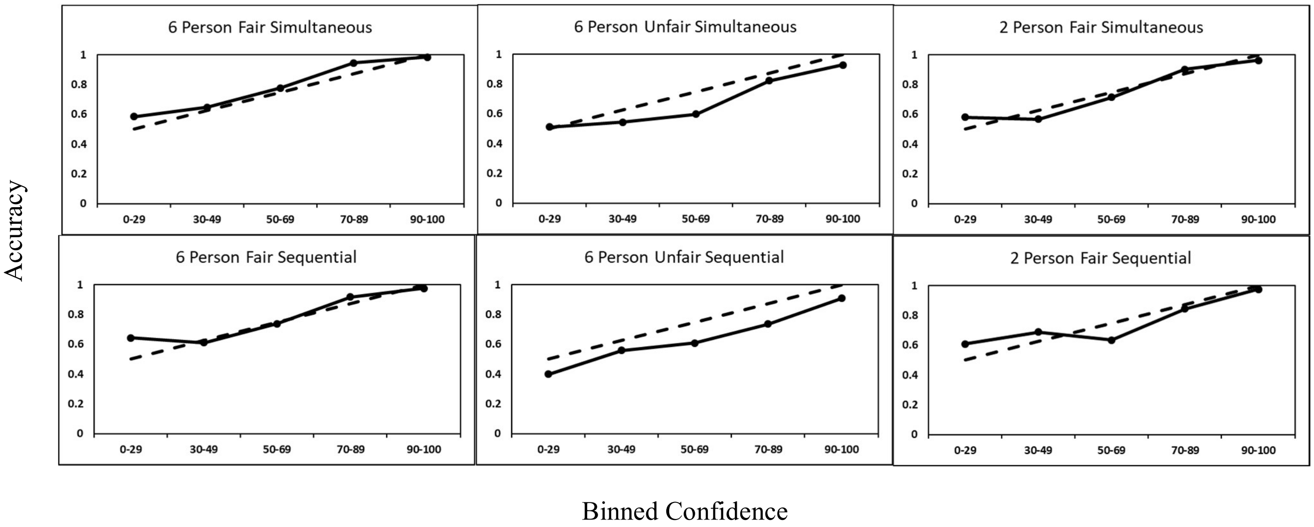


Figure 1. Confidence-accuracy characteristic curves for choosers in each of the six conditions. Binned confidence levels are displayed on the x-axis and accuracy given that an eyewitness identified a suspect is displayed on the y-axis. Dotted lines represent a perfectly calibrated confidence-accuracy relationship.

In all six conditions, confidence did predict accuracy: as confidence increased accuracy also increased. However, it is also apparent that accuracy was lower for the six-person unfair lineups, as the rate of correct identifications was always below the perfectly calibrated confidence accuracy relationship. These graphs do not appear to show any large differences between simultaneous lineups and sequential lineups nor any interactions between lineup fairness and lineup presentation style.

We also generated CAC curves for non-choosers (see Figure 2). In this case, we compared correct rejections of target absent lineups with incorrect rejections of target present lineups. The accuracy score represents the rate of correct rejections at a certain confidence level given that the eyewitness chose to reject the lineup. In this case, confidence does not predict accuracy for any of the six conditions, as there is no meaningful relationship between confidence and accuracy in the graphs. To summarize, confidence reliably predicted accuracy for choices, but not for non-choices.

Figure 2

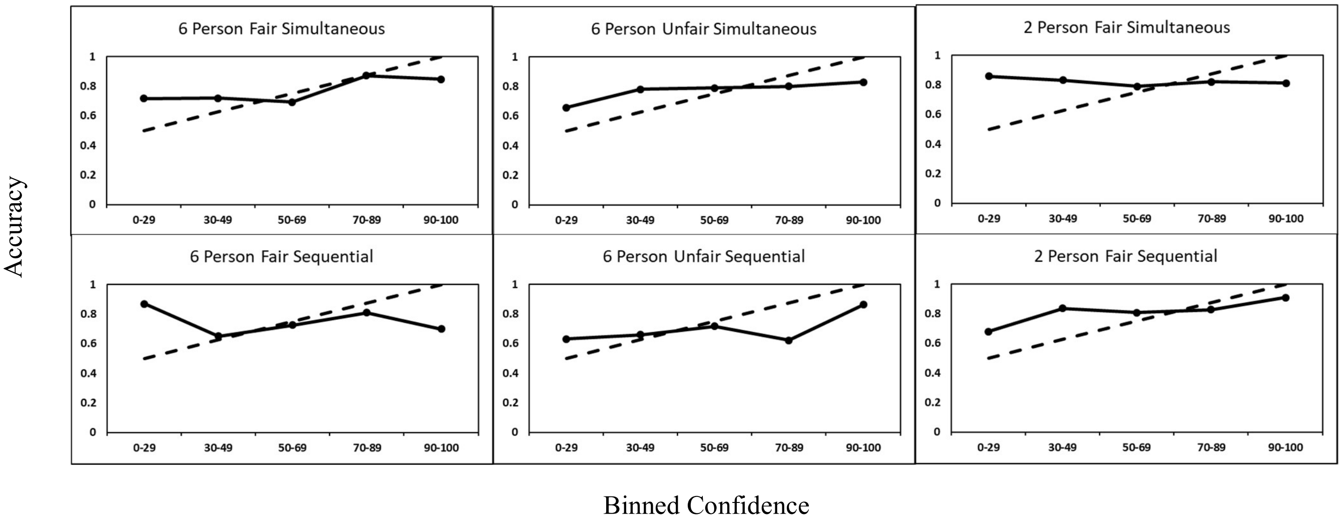


Figure 2. Confidence-accuracy characteristic curves for non-choosers in each of the six conditions. Binned confidence levels are displayed on the x-axis and accuracy given that an eyewitness rejected a lineup is displayed on the y-axis. Dotted lines represent a perfectly calibrated confidence-accuracy relationship.

Receiver operating characteristic curves. ROC curves have recently been used as a means to evaluate the difference in diagnosticity between multiple lineup conditions. While *d'* provides an overall diagnosticity score, an ROC curve provides a more complete picture of the diagnosticity of a lineup procedure as confidence score cutoffs become more conservative (see Wixted & Mickes, 2012). To create ROC curves, we used the same confidence bins that we did for the Somers' D and CAC analyses: 0-29, 30-49, 50-69, 70-89, 90-100. We then added up the number of correct identifications from target present lineups and the number of false identifications from target absent lineups for each confidence bin aggregated across participants. and divided the number of false identifications from target absent lineups by the number of plausible choices. The first point plotted on the graph (farthest from the axes) represents the rate of correct identifications and false alarms aggregated across confidence bins. For each subsequent point, we eliminated correct identifications and false alarms from the lowest confidence bin and recalculated the rates until the last point, which includes only the rates of correct identifications and false alarms for the 90-100 confidence bin.

The ROC curves for each of the conditions can be seen in Figure 3. The graphs show that the two-person fair lineups have the highest correct identification rates of the three fairness conditions and that the six-person unfair lineups have the highest false alarm rates. The six-person fair lineups have both lower correct identification and lower false alarm rates than the other conditions. It can also be seen that sequential lineups consistently have fewer correct identifications than simultaneous lineups, but that for the six-person unfair and two-person fair lineups the sequential lineups also have fewer false alarms, a finding that is consistent with previous studies (Dobolyi & Dodson, 2013; Gronlund et al., 2012; Mickes, Flowe, & Wixted, 2012).

Figure 3

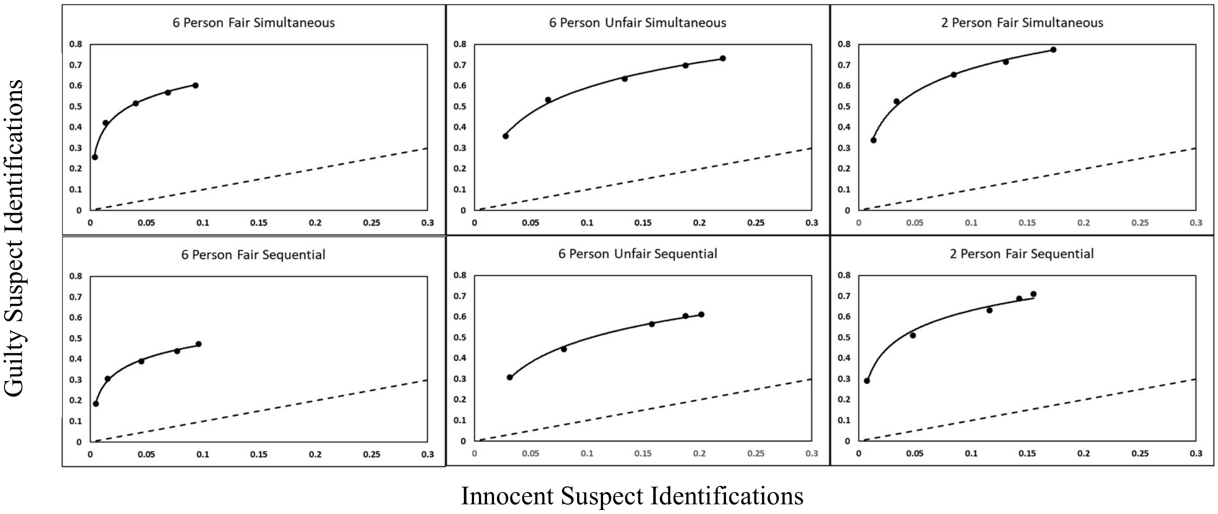


Figure 3. Receiver operating characteristic curves for choosers in each of the six conditions. False alarm rates from target absent lineups are displayed on the x-axis and correct identification rates from target present lineups are displayed on the y-axis. Dotted lines represent chance accuracy (equal rates of correct identifications and false alarms).

Traditionally, the method of comparing ROC curves has been to calculate and compare the partial area under the curve (pAUC) of each lineup. While this is an effective comparison for two lineup procedures with equal false alarm rates, it is problematic when the false alarm rates are different. This is because in order to make a comparison, you must either extend the length of the curve with fewer false alarms, thus projecting data which was never collected, or restrict the length of the curve with more false alarms, thus ignoring data (Smith, Lampinen, Wells, Smalarz, & Mackovichova, 2019). Smith et al. (2019) suggest that in order to address this, eyewitness researchers should instead use a methodology in which we measure each point's distance from perfect performance (i.e., when the correct identification rate is 1 and the false alarm rate is 0). This measure, termed deviation from perfect performance (*DPP*), is calculated by adding together the false alarm rate and one minus the correct identification rate for each confidence point. Greater *DPP* scores represent a larger deviance from perfect performance, so the best lineup procedure will have the smallest *DPP* value.

Table 6 displays the correct identification rates, false alarm rates, and *DPP* values for each of the six conditions. It can be seen here that in all cases the simultaneous lineups are superior to the sequential lineups. It is also the case that the two-person fair lineups are superior to the six-person fair and six-person unfair lineups. For all confidence levels except the 90-100 confidence bin, the simultaneous two-person fair lineup has the lowest *DPP* score among the six conditions. For confidence bin 90-100, however, the simultaneous six-person unfair lineup has the lowest *DPP* score. Consistent with our other analyses, these results do not suggest that there is an interaction between lineup fairness condition and lineup presentation and in fact the differences between the *DPP* scores for the three fairness conditions are actually larger for sequential lineups than simultaneous lineups, which is the opposite of what we predicted. These results also show that *DPP* scores are lowest when all confidence levels are included in the analysis (i.e., the 0-29 confidence level in Table 6) in all conditions except the sequential six-person unfair lineup.

It is important, however, to acknowledge the assumptions at play in *DPP* analysis. In using the simplest *DPP* model we are assuming that correct identifications of a guilty suspect are as good as false identifications of an innocent suspect are bad. As Ceci and Friedman (2000) note according to the United States constitution a false conviction is considered a worse outcome than the failure to convict. If we readjust the weights between correct identifications and false identifications, we might instead come to the conclusion that the six-person fair lineups are superior, as they produce both fewer correct identifications and fewer false identifications.

Table 6

DPP Values by Lineup Presentation and Lineup Fairness Conditions

	6 Person Fair					
	Simultaneous			***Sequential***		
Confidence	**Correct ID**	**False Alarm**	***DPP***	**Correct ID**	**False Alarm**	***DPP***
90-100	0.256	0.004	0.748	0.186	0.005	0.818
70-89	0.422	0.014	0.592	0.305	0.016	0.710
50-69	0.515	0.040	0.526	0.390	0.046	0.656
30-49	0.567	0.069	0.502	0.439	0.077	0.638
0-29	0.602	0.094	0.492	0.474	0.096	0.623
		Average *DPP* =	**0.572**		**Average *DPP* =**	**0.689**

	6 Person Unfair					
	Simultaneous			***Sequential***		
Confidence	**Correct ID**	**False Alarm**	***DPP***	**Correct ID**	**False Alarm**	***DPP***
90-100	0.358	0.028	0.670	0.309	0.031	0.722
70-89	0.532	0.065	0.533	0.444	0.080	0.636
50-69	0.634	0.134	0.500	0.566	0.158	0.592
30-49	0.698	0.188	0.490	0.603	0.188	0.584
0-29	0.733	0.221	0.488	0.613	0.202	0.589
		Average *DPP* =	**0.536**		**Average *DPP* =**	**0.625**

	2 Person Fair					
	Simultaneous			***Sequential***		
Confidence	**Correct ID**	**False Alarm**	***DPP***	**Correct ID**	**False Alarm**	***DPP***
90-100	0.337	0.013	0.676	0.291	0.007	0.717
70-89	0.526	0.033	0.507	0.511	0.048	0.536
50-69	0.654	0.084	0.430	0.631	0.116	0.485
30-49	0.715	0.131	0.416	0.689	0.142	0.453
0-29	0.773	0.173	0.400	0.709	0.156	0.446
		Average *DPP* =	**0.486**		**Average *DPP* =**	**0.528**

CHAPTER 4

DISCUSSION AND CONCLUSION

The current study was conducted for a few main purposes. Firstly, we were interested in replicating previous findings that eyewitness accuracy would be higher for pristine as opposed to non-pristine lineup conditions. The results confirm our hypothesis that accuracy would be higher for fair lineups than for unfair lineups, at least when we control for difficulty by comparing lineups which have the same number of plausible choices. This is consistent with the findings of Charman et al. (2011) and also extends these findings to sequential lineups. Secondly, we were interested in replicating additional findings that confidence in false alarms would be greater for non-pristine conditions than for pristine conditions. The results confirm the prediction that confidence in false alarms would be higher for unfair lineups as opposed to fair lineups, which also is consistent with findings from Charman et al. (2011). Thirdly, we were interested in determining whether the difference in the confidence-accuracy relationship would be greater for simultaneous lineups than for sequential lineups. We predicted that the difference in the confidence-accuracy relationship between pristine and non-pristine conditions could be explained by differential usage of absolute versus relative judgments and that therefore we could reduce this difference by using sequential lineups by restricting participants towards using absolute judgments more as opposed to relative judgments. The results did not support this hypothesis. None of the analyses analyzing accuracy, choosing rate, confidence, or confidence and accuracy together produced a significant interaction between lineup fairness and lineup presentation style. Possible explanations for these unexpected findings are considered here.

One explanation for why we did not see any significant interactions between lineup fairness and lineup presentation style is that differential use of absolute versus relative judgments may not meaningfully affect confidence and accuracy or their relationship. Weber and Brewer (2004) previously found that there was no significant difference in the confidence-accuracy relationship between simultaneous and sequential lineups and our findings add to this by suggesting that lineup fairness does not differentially affect the confidence-accuracy relationship for simultaneous versus sequential lineups. However, another possible explanation for these findings is that the simultaneous and sequential lineup distinction may not be sufficiently manipulating absolute versus relative judgments. The fact that we found that simply adding dissimilar foils to a sequential lineup significantly decreased accuracy suggests that eyewitnesses are making relative judgments even in sequential lineups. Future research should consider other measures of absolute versus relative judgment. Self-report may be a useful measure to consider. For example, Dunning and Stern (1994) found that eyewitnesses who self-reported using absolute judgments were more accurate than eyewitnesses who self-reported using relative judgments. Eye-tracking can also provide additional insight into eyewitnesses' use of absolute versus relative judgments (e.g., Starns et al., 2017).

This study also provides evidence relating to the current debate between use of simultaneous and sequential lineups. While sequential lineups have been recommended as superior to simultaneous lineups in the past (Wells et al., 1998), it has been argued more recently that switching to sequential lineups involves a trade-off: decreased discriminability in favor of more conservative choosing rates (Gronlund, Mickes, Wixted, & Clark, 2016). Our results are consistent with this understanding of the trade-off between simultaneous and sequential lineups. We also found that eyewitnesses were significantly more confident in false alarms from sequential lineups as compared to simultaneous lineups, which is consistent with previous research (Dobolyi & Dodson, 2013) and suggests that sequential lineups may not necessarily be superior to simultaneous lineups.

We additionally considered the role of estimator variables in the eyewitness process. While we did not replicate the previous findings of Andersen et al. (2014) which suggest that working memory capacity predicts eyewitness identification accuracy, we did find novel results about the role of suggestibility in eyewitness identification. The GSS 1 has been commonly used in forensic settings as a measure to predict susceptibility to suggestion during interrogation; however, our study provides new information by examining its predictive value in eyewitness identification procedures. Specifically, we found that individual differences in suggestibility robustly predicted eyewitness choosing rates, with more suggestible eyewitnesses making more liberal choices from lineups. In real world lineups there is often a great deal of suggestion, as it is easy to assume that if you are presented with a lineup by a police officer that they must believe that they have found the guilty suspect. The fact that we found that the GSS 1 significantly predicted choosing rates and *d'* scores in a laboratory study suggests that this may be a useful tool for assessing eyewitness identification accuracy.

Future Directions

One clear limitation of the current study is that we are aiming to describe the general distinction in the confidence-accuracy relationship between pristine and non-pristine conditions, but we only manipulated the single pristineness condition that the suspect should not stand out in the lineup. It is possible that other pristineness manipulations may differentially affect use of absolute versus relative judgments in identification. Therefore, we are currently running a follow-up study to address a second of Wixted and Wells' (2017) prescriptions for a robust confidence-accuracy relationship: that the confidence statement should be taken at the time of the first identification. In real world settings it is often the case that eyewitnesses must make identifications with the same suspect multiple times throughout an investigation (Behrman & Davey, 2001). In fact, in some cases multiple identifications are required by law. For example, in England and Wales if a live showup is disputed, a subsequent lineup identification must occur (Valentine, Davis, Memon, & Roberts, 2012).

Conducting multiple identification procedures for the same suspect can be problematic. One reason for this is that eyewitnesses who complete multiple identification procedures may be exposed to post-identification feedback. In a large body of literature, positive feedback was found to significantly increase eyewitness confidence in subsequent identification procedures (see Steblay, Wells, & Douglass, 2014, for a meta-analysis). However, even in cases in which no post-identification feedback is provided, there are still concerns about conducting multiple identification procedures for the same suspect. In fact, exposure to an innocent suspect in a mugshot (Deffenbacher, Bornstein, & Penrod, 2006), in a showup (Godfrey & Clark, 2010; Haw, Dickinson, & Meissner, 2007; Lawson & Dysart, 2014) or in a lineup (Hinz & Pezdek, 2001; Pezdek & Blandon-Gitlin, 2005) has been found to increase false identifications of that suspect in subsequent lineups. Prior eyewitness identification procedures have also been found to affect eyewitness confidence (e.g., Godfrey and Clark, 2010). Overall, the literature suggests that both accuracy and the relationship between accuracy and confidence are compromised by repeated testing. Analogous to the current study, we are currently exploring whether this distinction can be explained by differential use of absolute and relative judgments.

A summary and preliminary analysis of our follow-up study is presented here. The methodology of this study is similar to the current study. Thirty-two participants (20 female, 12 male, M age = 20.4 years, SD = 1.64), 40% of the planned sample size, have completed the study to date. Each participant was first presented with 24 videos that were used in the current study (half male and half female). After completing a filler task (operation span task), participants made identifications from 12 lineups each corresponding to one of the 24 videos. After a two-day delay, participants then completed 24 lineups, 12 of which they had seen in the immediate test and 12 of which were new. Participants were randomly assigned to receive either simultaneous or sequential lineups and each participant received half target present and half target absent lineups.

We predicted that confidence in incorrect identifications would be higher for repeated identifications than for first-time identifications and that confidence and accuracy would be better calibrated for first-time identifications than for second-time identifications. Additionally, we predicted that the difference in the confidence-accuracy relationship between first-time and second-time identifications would be greater for simultaneous lineups than for sequential lineups.

We include preliminary data for the accuracy, choosing rates, confidence levels, and chooser Somers' D rates of this study. Table 7 displays the rates at which participants made each of the five identification decisions (correct identification, foil identification, incorrect rejection, correct rejection, and false alarm). As can be seen, the repeated lineups had the highest rates of false identifications from target absent lineups, but the delayed lineups had the lowest rates of correct identifications from target present lineups.

Table 7

Identification Performance by Lineup Presentation and Lineup Time/Repetition Conditions

		Target Present Lineups						**Target Absent Lineups**			
		Correct Identifications		**Foil Identifications**		**Incorrect Rejections**		**Correct Rejections**		**False Alarms**	
Presentation	**Time/Repetition**	**M**	**SD**	**M**	**SD**	**M**	**SD**	**M**	**SD**	**M**	**SD**
Simultaneous	Immediate	0.64	0.25	0.19	0.20	0.18	0.22	0.42	0.19	0.58	0.19
Simultaneous	Repeat	0.66	0.31	0.23	0.23	0.11	0.15	0.27	0.21	0.73	0.21
Simultaneous	Delay	0.49	0.27	0.26	0.23	0.25	0.19	0.62	0.31	0.38	0.31
Sequential	Immediate	0.63	0.22	0.24	0.22	0.13	0.13	0.56	0.32	0.44	0.32
Sequential	Repeat	0.48	0.25	0.43	0.26	0.10	0.13	0.20	0.22	0.80	0.22
Sequential	Delay	0.39	0.21	0.27	0.29	0.33	0.24	0.70	0.35	0.30	0.35

Preliminary data on the accuracy measures, *d’* and P_r, can be seen in Table 8. *d’* results show a clear pattern of accuracy being highest for immediate identifications and lowest for repeated identifications. The results, however, currently show an interaction pattern in the opposite direction of our predictions: the difference in the accuracy for the different pristineness conditions is larger for the sequential lineups than for the simultaneous lineups. The P_r results suggest that immediate and repeated testing are no different for simultaneous lineups, but that they produce higher accuracy than delayed lineups. In contrast, the results for sequential lineups suggest that accuracy is higher for immediate lineups than for either repeated or delayed lineups.

Table 8

Accuracy Measures by Lineup Presentation and Lineup Time/Repetition Conditions

		d’		P_r	
Presentation	**Time/Repetition**	**M**	**SD**	**M**	**SD**
Simultaneous	Immediate	1.83	0.94	0.54	0.25
Simultaneous	Repeat	1.71	1.10	0.54	0.30
Simultaneous	Delay	1.75	1.19	0.43	0.28
Sequential	Immediate	2.08	1.24	0.56	0.25
Sequential	Repeat	1.00	0.85	0.34	0.25
Sequential	Delay	1.56	1.18	0.34	0.25

Preliminary data on the choosing rate measures, *C* and B_r, can be found in Table 9. Both measures show that eyewitnesses are more conservative in their identifications from delayed lineups than either immediate lineups or repeated lineups. Both measures also show that eyewitnesses are more conservative in their identifications from sequential lineups than from simultaneous lineups.

Table 9

Choosing Rate Measures by Lineup Presentation and Lineup Time/Repetition Conditions

		C		***B_r***	
Presentation	**Time/Repetition**	**M**	**SD**	**M**	**SD**
Simultaneous	Immediate	0.40	0.45	0.32	0.26
Simultaneous	Repeat	0.34	0.60	0.38	0.26
Simultaneous	Delay	0.90	0.55	0.13	0.09
Sequential	Immediate	0.62	0.28	0.16	0.10
Sequential	Repeat	0.64	0.41	0.22	0.08
Sequential	Delay	1.13	0.34	0.07	0.06

Confidence data is displayed in Table 10. Consistent with the current study, confidence is higher for sequential lineups than for simultaneous lineups overall. In aggregate, it is also apparent that confidence is greater for immediate lineups than for repeated lineups or delayed lineups.

Table 10

Confidence by Lineup Presentation and Lineup Time/Repetition Conditions

		Target Present Lineups						**Target Absent Lineups**			
		Correct Identifications		**Foil Identifications**		**Incorrect Rejections**		**Correct Rejections**		**False Alarms**	
Presentation	**Time/Repetition**	**M**	**SD**	**M**	**SD**	**M**	**SD**	**M**	**SD**	**M**	**SD**
Simultaneous	Immediate	77.7	18.4	51.8	29.8	35.7	23.4	43.1	20.5	46.3	12.1
Simultaneous	Repeat	73.6	17.1	44.9	29.1	30.7	24.4	33.2	21.6	46.0	19.4
Simultaneous	Delay	62.0	27.0	32.3	22.9	33.2	22.9	36.0	22.8	35.0	18.8
Sequential	Immediate	75.4	18.0	53.0	21.3	49.1	19.6	44.7	21.3	54.0	17.9
Sequential	Repeat	74.6	26.9	54.6	26.0	35.2	25.2	38.6	21.6	48.5	21.4
Sequential	Delay	66.9	25.0	50.2	23.3	39.2	25.8	38.1	28.0	43.2	20.6

Somers' D results are displayed in Table 11. The results suggest that the confidence-accuracy relationship is stronger for sequential lineups than simultaneous lineups overall. For both simultaneous and sequential lineups, Somers' D scores are higher for lineups presented for the first time at a delay than for lineups than have been presented previously. The difference is currently larger for sequential than for simultaneous lineups, the opposite of what we predicted.

Table 11

Chooser Somers' D Values by Lineup Presentation and Lineup Time/Repetition Conditions

Presentation	**Time/Repetition**	**M**	**SD**
Simultaneous	Immediate	0.54	0.41
Simultaneous	Repeat	0.40	0.46
Simultaneous	Delay	0.44	0.40
Sequential	Immediate	0.51	0.38
Sequential	Repeat	0.56	0.27
Sequential	Delay	0.72	0.21

Overall, it is too early to make conclusive statements about the results of this study, however, it is already clear that accuracy and confidence are highest for lineups conducted immediately after viewing the crime videos. Additionally, the results suggest that false alarms are more likely to occur with repeated testing, but that the lowest rates of correct identifications are from lineups that were presented for the first time at the delay.

Conclusion

The aim of our study was to better understand what differentiates the conditions under which eyewitness confidence predicts accuracy from the conditions under which it may be less predictive. Based on the task calibration principle of FTT, we predicted that the difference in the accuracy-confidence relationship could be predicted by differential use of absolute and relative judgments in the categorical decision of making an identification from a lineup. While our findings do not give us a definitive answer to this question, they do give us insight into the use of simultaneous and sequential lineups. Specifically, we find that the addition of visually dissimilar foils to a lineup decreases eyewitness accuracy and increases confidence in false identifications, regardless of whether the lineup is conducted simultaneously or sequentially. This suggests that eyewitnesses are making relative judgments in sequential lineups. Therefore, we can make the argument that the distinction between simultaneous and sequential lineups may not be a useful manipulation of absolute versus relative judgments. While we already know that certain testing conditions better support a strong confidence-accuracy relationship than others, this line of research's implications are knowledge of the theoretical distinctions between pristine and non-pristine conditions. By gaining a fuller understanding of this distinction, we can better equip our legal institutions to promote practices which will provide more informative eyewitness evidence.

References

Andersen, S. M., Carlson, C. A., Carlson, M. A., & Gronlund, S. D. (2014). Individual differences predict eyewitness identification performance. *Personality and Individual Differences, 60*, 36-40.

Behrman, B. W., & Davey, S. L. (2001). Eyewitness identification in actual criminal cases: An archival analysis. *Law and Human Behavior, 25*, 475-491.

Brainerd, C. J., Nakamura, K., Reyna, V. F., & Holliday, R. E. (2017). Overdistribution illusions: Categorical judgments produce them, confidence ratings reduce them. *Journal of Experimental Psychology: General, 146*, 20-40.

Brainerd, C. J., & Reyna, V. F. (2005). *The science of false memory*. New York: Cambridge University Press.

Ceci, S. J., & Friedman, R. D. (2000). The suggestibility of children: Scientific research and legal implications. *Cornell Law Review, 86*, 33-108.

Charman, S. D., Wells, G. L., & Joy, S. W. (2011). The dud effect: Adding highly dissimilar fillers increases confidence in lineup identifications. *Law and Human Behavior, 35*, 479-500.

Corbin, J. C., Reyna, V. F., Weldon, R. B., & Brainerd, C. J. (2015). How reasoning, judgment, and decision making are colored by gist-based intuition: A fuzzy-trace theory approach. *Journal of Applied Research in Memory and Cognition, 4*, 344-355.

Cutler, B. L., Penrod, S. D., & Stuve, T. E. (1988). Juror decision making in eyewitness identification cases. *Law and Human Behavior, 12*, 41-55.

Darling, S., Martin, D., Hellmann, J. H., & Memon, A. (2009). Some witnesses are better than others. *Personality and Individual Differences, 47*, 369-373.

Deffenbacher, K. A., Bornstein, B. H., & Penrod, S. D. (2006). Mugshot exposure effects: Retroactive interference, mugshot commitment, source confusion, and unconscious transference. *Law and Human Behavior, 30*, 287-307.

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

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

Dunning, D., & Stern, L. B. (1994). Distinguishing accurate from inaccurate eyewitness identifications via inquiries about decision processes. *Journal of Personality and Social Psychology, 67*, 818-835.

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

Godfrey, R. D., & Clark, S. E. (2010). Repeated eyewitness identification procedures: Memory, decision making, and probative value. *Law and Human Behavior, 34*, 241-258.

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

Gronlund, S. D., Mickes, L., Wixted, J. T., & Clark, S. E. (2015). Conducting an eyewitness lineup: How the research got it wrong. In B. H. Ross (Ed.), *The psychology of learning and motivation: Vol. 63* (pp. 1-43). Waltham, MA: Academic Press.

Gudjonsson, G. H. (1984a). A new scale of interrogative suggestibility. *Personality and Individual Differences, 5*, 303-314.

Gudjonsson, G. H. (1984b). Interrogative suggestibility: Comparison between 'false confessors' and 'deniers' in criminal trials. *Medicine, Science and the Law, 24*, 56-60.

Hanczakowski, M., Zawadzka, K., & Higham, P. A. (2014). The dud-alternative effect in memory for associations: Putting confidence into local context. *Psychonomic Bulletin & Review, 21*, 543-548.

Haw, R. M., Dickinson, J. J., & Meissner, C. A. (2007). The phenomenology of carryover effects between show-up and line-up identification. *Memory, 16*, 117-127.

Hinz, T., & Pezdek, K. (2001). The effect of exposure to multiple lineups on face identification accuracy. *Law and Human Behavior, 25*, 185-198.

Innocence Project (2018). DNA Exonerations in the United States. Retrieved May 2, 2019. https://www.innocenceproject.org/dna-exonerations-in-the-united-states/

Lawson, V. Z., & Dysart, J. E. (2014). The showup identification procedure: An exploration of systematic bias. *Legal and Criminological Psychology, 19*, 54-68.

Lindsay, R. C. L., & Wells, G. L. (1980). What price justice? Exploring the relationship of lineup fairness to identification accuracy. *Law and Human Behavior, 4*, 303-313.

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

Mansour, J. K., Beaudry, J. L., Bertrand, M. I., Kalmet, N., Melsom, E. I., & Lindsay, R. C. L. (2012). Impact of disguise on identification decisions and confidence with simultaneous and sequential lineups. *Law and Human Behavior, 36*, 513-526.

Merckelbach, H., Muris, P., Wessel, I., & Van Koppen, P. J. (1998). The Gudjonsson Suggestibility Scale (GSS): Further data on its reliability, validity, and metacognition correlates. *Social Behavior and Personality, 26*, 203-210.

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

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.

Morgan, C. A., III, Hazlett, G., Baranoski, M., Doran, A., Southwick, S., & Loftus, E. (2007). Accuracy of eyewitness identification is significantly associated with performance on a standardized test of face recognition. *International Journal of Law and Psychiatry, 30*, 213-223.

Pannu, J. K., & Kaszniak, A. W. (2005). Metamemory experiments in neurological populations: A review. *Neuropsychology Review, 15*, 105-130.

Pezdek, K., & Blandon-Gitlin, I. (2005). When is an intervening line-up most likely to affect eyewitness identification accuracy? *Legal and Criminological Psychology, 10*, 247-263.

Roediger, H. L., III, & DeSoto, K. A. (2014). Confidence and memory: Assessing positive and negative correlations. *Memory, 22*, 76-91.

Seamon, J. G., Luo, C. R., Kopecky, J. J., Price, C. A., Rothschild, L., Fung, N. S., & Schwartz, M. A. (2002). Are false memories more difficult to forget than accurate memories?: The effect of retention interval on recall and recognition. *Memory & Cognition, 30,* 1054-1064.

Searcy, J. H., Bartlett, J. C., & Memon, A. (1999). Age differences in accuracy and choosing in eyewitness identification and face recognition. *Memory & Cognition, 27*, 538-552.

Selmeczy, D., & Dobbins, I. G. (2014). Relating the content and confidence of recognition judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*, 66-85.

Smith, A. M., Lampinen, J. M., Wells, G. L., Smalarz, L., & Mackovichova, S. (2019). Deviation from perfect performance measures the diagnostic utility of eyewitness lineups but partial area under the ROC curve does not. *Journal of Applied Research in Memory and Cognition, 8*, 50-59.

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General, 117*, 34-50.

Somers, R. H. (1962). A new asymmetric measure of association for ordinal variables. *American Sociological Review, 27*, 799-811.

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers, 31*, 137-149.

Starns, J. J., Chen, T., & Staub, A. (2017). Eye movements in forced-choice recognition: Absolute judgments can preclude relative judgments. *Journal of Memory and Language, 93*, 55-66.

Steblay, N. K., Wells, G. L., & Douglass, A. B. (2014). The eyewitness post-identification feedback effect 15 years later: Theoretical and policy implications. *Psychology, Public Policy, and Law, 20*, 1-18.

Valentine, T., Davis, J. P., Memon, A., & Roberts, A. (2012). Live showups and their influence on subsequent video line-up. *Applied Cognitive Psychology, 26*, 1-23.

Von der Malsburg, T. (2015). Py-span-task – A software for testing working memory span.

Weber, N., & Brewer, N. (2004). Confidence-accuracy calibration in absolute and relative face recognition judgments. *Journal of Experimental Psychology: Applied, 10*, 156-172.

Wells, G. L. (1984). The psychology of lineup identification. *Journal of Applied Social Psychology, 14*, 89-103.

Wells, G. L., Leippe, M. R., & Ostrom, T. M. (1979). Guidelines for empirically assessing the fairness of a lineup. *Law and Human Behavior, 3*, 285-293.

Wells, G. L., Rydell, S. M., & Seelau, E. P. (1993). The selection of distractors for eyewitness lineups. *Journal of Applied Psychology, 78*, 835-844.

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

Windschitl, P. D., & Chambers, J. R. (2004). The dud-alternative effect in likelihood judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*, 198-215.

Wixted, J. T., & Mickes, L. (2012). The field of eyewitness memory should abandon probative value and embrace receiver operating characteristic analysis. *Perspectives on Psychological Science, 7*, 275-278.

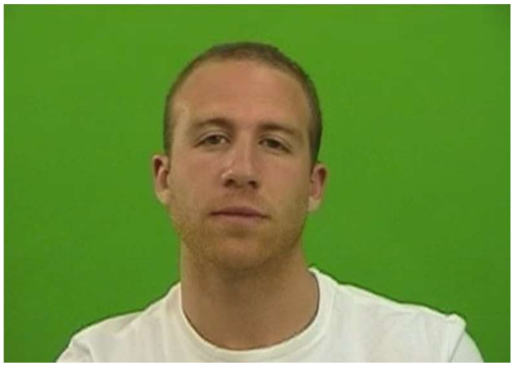
Wixted, J. T., & Wells, G. L. (2017). The relationship between eyewitness confidence and identification accuracy: A new synthesis. *Psychological Science in the Public Interest, 18*, 10-65.

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*, 552-564.

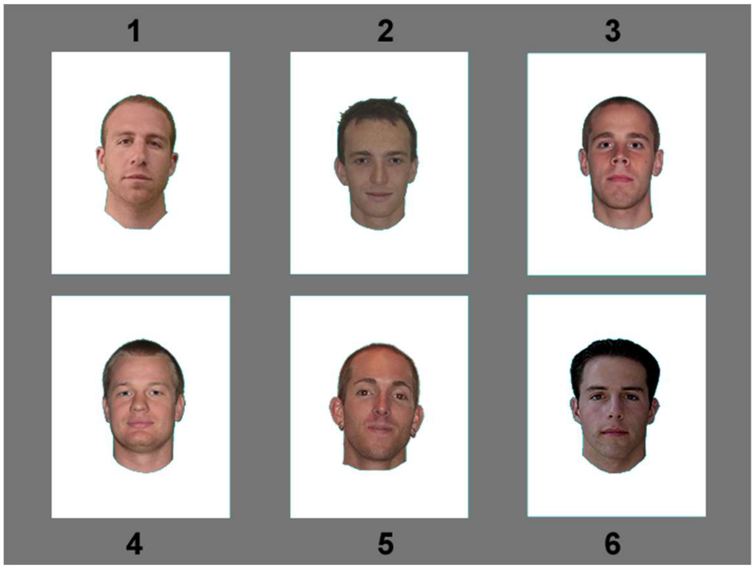
Appendix A

Lineup Examples

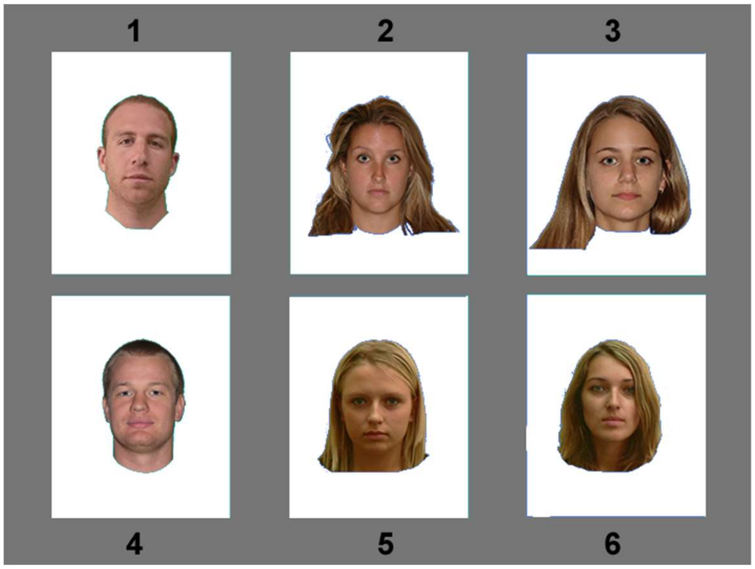
A. Single Frame of Video



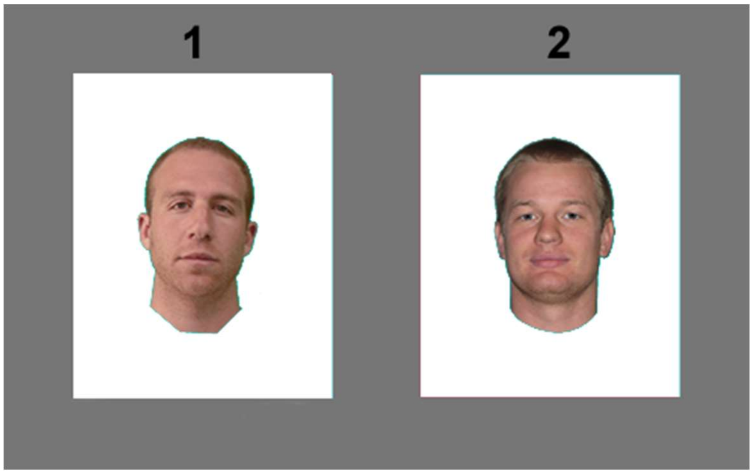
B. Target Present Six-Person Fair Lineup



C. Target Present Six-Person Unfair Lineup



D. Target Present Two-Person Fair Lineup



Appendix B

Gudjonsson Materials

A. Comparison between original Gudjonsson transcripts and American adaption

Original GSS1 transcript. Words in bold were replaced in the American adaption.

> Anna Thomson of South **Croydon** was on **holiday** in Spain when she was held up outside her hotel and robbed of her handbag which contained **£**50 worth of travellers cheques and her passport. She screamed for help and attempted to put up a fight by kicking one of the assailants in the shins. A police car shortly arrived and the woman was taken to the nearest police station where she was interviewed by Detective Sergeant Delgado. The woman reported that she had been attacked by three men one of whom she described as **oriental** looking. The men were said to be slim and in their early twenties. The police officer was touched by the woman's story and advised her to contact the **British** Embassy. Six days later the police recovered the lady's handbag, but the contents were never found. Three men were subsequently charged two of whom were convicted and given prison sentences. Only one had had previous convictions for similar offences. The lady returned to **Britain** with her husband Simon and two friends but remained frightened of being out on her own.

American adaption of GSS1. Words in bold took the place of words in the original British version.

> Anna Thomson of South **Dakota** was on **vacation** in Spain when she was held up outside her hotel and robbed of her handbag which contained **$**50 worth of travellers cheques and her passport. She screamed for help and attempted to put up a fight by kicking one of the assailants in the shins. A police car shortly arrived and the woman was taken to the nearest police station where she was interviewed by Detective Sergeant Delgado. The woman reported that she had been attacked by three men one of whom she described as **Asian** looking. The men were said to be slim and in their early twenties. The police officer was touched by the woman's story and advised her to contact the **American** Embassy. Six days later the police recovered the lady's handbag, but the contents were never found. Three men were subsequently charged two of whom were convicted and given prison sentences. Only one had had previous convictions for similar offences. The lady returned to **the U.S.** with her husband Simon and two friends but remained frightened of being out on her own.

Original and adapted interview questions for GSS1. Bold words were changed in the adaption.

Original Questions	Adapted Questions
1. Did the woman have a husband called Simon? (NS)	1. Did the woman have a husband called Simon? (NS)
2. Did the woman have one or two children? (S)	2. Did the woman have one or two children? (S)
3. Did the woman's glasses break in the struggle? (S)	3. Did the woman's glasses break in the struggle? (S)
4. Was the woman's name Anna Wilkinson? (S)	4. Was the woman's name Anna Wilkinson? (S)
5. Was the woman interviewed by a detective sergeant? (NS)	5. Was the woman interviewed by a detective sergeant? (NS)
6. Were the assailants black or white? (S)	6. Were the assailants black or white? (S)
7. Was the woman taken to the central police station? (S)	7. Was the woman taken to the central police station? (S)
8. Did the woman's handbag get damaged in the struggle? (S)	8. Did the woman's handbag get damaged in the struggle? (S)
9. Was the woman on **holiday** in Spain? (NS)	9. Was the woman on **vacation** in Spain? (NS)
10. Were the assailants convicted six weeks after their arrest? (S)	10. Were the assailants convicted six weeks after their arrest? (S)
11. Did the woman's husband support her during the police interview? (S)	11. Did the woman's husband support her during the police interview? (S)
12. Did the woman hit one of the assailants with her fist or handbag? (S)	12. Did the woman hit one of the assailants with her fist or handbag? (S)
13. Was the woman from South **Croydon**? (NS)	13. Was the woman from South **Dakota**? (NS)
14. Did one of the assailants shout at the woman? (S)	14. Did one of the assailants shout at the woman? (S)
15. Were the assailants tall or short? (S)	15. Were the assailants tall or short? (S)
16. Did the woman's screams frighten the assailants? (S)	16. Did the woman's screams frighten the assailants? (S)
17. Was the police officer's name Delgado? (NS)	17. Was the police officer's name Delgado? (NS)
18. Did the police give the woman a lift back to her hotel? (S)	18. Did the police give the woman a lift back to her hotel? (S)
19. Were the assailants armed with knives or guns? (S)	19. Were the assailants armed with knives or guns? (S)
20. Did the woman's clothes get torn in the struggle? (S)	20. Did the woman's clothes get torn in the struggle? (S)

S = Suggestive questions
NS = Non-suggestive questions