Face Matching Lineups: An Application of the Filler-Control Method for Identification from Video Evidence

by

Camryn N. Yuen

Bachelor of Arts (Hons., Psychology), Simon Fraser University, 2021

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Arts

in the
Department of Psychology
Faculty of Arts and Social Sciences

© Camryn N. Yuen 2024
SIMON FRASER UNIVERSITY
Summer 2024

Copyright in this work is held by the author. Please ensure that any reproduction or re-use is done in accordance with the relevant national copyright legislation.

Declaration of Committee

Name:	**Camryn N. Yuen**
Degree:	**Master of Arts**
Title:	**Face Matching Lineups: An Application of the Filler-Control Method for Identification from Video Evidence**
Committee:	**Chair: Ralph Mistlberger** Professor, Psychology **Ryan Fitzgerald** Supervisor Associate Professor, Psychology **Daniel Bernstein** Committee Member Adjunct Faculty, Psychology **Kara Moore** Examiner Assistant Professor, Psychology University of Utah

Ethics Statement

The author, whose name appears on the title page of this work, has obtained, for the research described in this work, either:

a. human research ethics approval from the Simon Fraser University Office of Research Ethics

or

b. advance approval of the animal care protocol from the University Animal Care Committee of Simon Fraser University

or has conducted the research

c. as a co-investigator, collaborator, or research assistant in a research project approved in advance.

A copy of the approval letter has been filed with the Theses Office of the University Library at the time of submission of this thesis or project.

The original application for approval and letter of approval are filed with the relevant offices. Inquiries may be directed to those authorities.

Simon Fraser University Library
Burnaby, British Columbia, Canada

Update Spring 2016

Abstract

When a crime is caught on camera, investigators may use face matching to identify a suspect with a physical appearance that matches the perpetrator in the footage. Typically, the identification is from a 1-to-1 comparison between the video evidence and a suspect image. The purpose of this thesis was to test decision-making in face matching lineups, which included the suspect's image alongside images of five non-suspects. Participants made a series of face matching decisions, using the lineup or the 1-to-1 procedure. Face matching lineups were able to improve suspect identification accuracy over the 1-to-1 procedure. However, this was only true if it was assumed that the lineups were fair and that the risk to an innocent suspect was equal to chance. Although further research is needed on how achieve this level of fairness, face matching lineups could be a promising solution to the risk of misidentification from video evidence.

Keywords: face matching; video evidence identification; 1-to1 comparison; face matching lineups

This thesis is dedicated to all who have supported me, including my supervisor, my parents, my partner, and my friends.

Table of Contents

Declaration of Committee ii
Ethics Statement iii
Abstract iv
Dedication v
Table of Contents vi
List of Tables viii
List of Figures viii
Glossary ix

Introduction 1
Face Matching 2
The Filler-Control Method 6
Current Study 11
Hypothesis 12

Method 13
Participants 13
Design 13
Materials 14
Crime Video Images and Perpetrator Mugshots 14
Lineup Condition 15
1-to-1 Condition 16
Practice Trials 17
Procedure 17

Data Analysis 20
Accuracy 20
Positive predictive value 21
Negative predictive value 21
Receiver operating characteristic (ROC) analysis 21
1-to-1 Task 22
Lineup 22
Rule-out procedure 23
Confidence-accuracy characteristic (CAC) analysis 23

Results 25
Preregistered Hypothesis Test 25
Exploratory Analyses 25
Correct Identifications 25
Positive Predictive Value 25
PPV, assuming all-suspect lineup 25
PPV, assuming innocent suspect is among the plausible lineup options 26
PPV, assuming perfect lineup fairness 26

Negative Predictive Value 26
Receiver operating characteristic (ROC) curves 27
Area under the curve (AUC) 28
Confidence-accuracy characteristic (CAC) curves 28

Discussion 31
Limitations...37
Future Directions...38
Conclusion 39

References...40
Cases..50

List of Tables

Table 1 Definition of key terms 7
Table 2 Counterbalancing of 1-to-1 and lineup trials 17
Table 3 Face matching accuracy rate for 1-to-1 and lineup conditions 27

List of Figures

Figure 1 Matching crime video and mugshot image 15
Figure 2 Mismatching crime video and mugshot image 15
Figure 3 Receiver operating characteristic curves 27
Figure 4 Confidence-accuracy characteristic curves: Positive predictive value 29
Figure 5 Confidence-accuracy characteristic curves: Negative predictive value 30

Glossary

Effective size correction	A method for estimating innocent suspect identifications. Involves dividing the overall false alarm rate in target-absent lineups by effective size (i.e., number of plausible lineup members). A lineup member is plausible if they attract a certain number of identifications.
Filler-control method	Presenting a suspect alongside a group of known-innocent fillers.
Fillers	Lineup members who are known to be innocent.
Negative predictive value (NPV)	The proportion of rejections that are correct. NPV is calculated by dividing correct rejections by the sum of correct rejections and incorrect rejections.
Nominal size correction	A method for estimating innocent suspect identifications. Involves dividing the overall false alarm rate in target-absent lineups by nominal size (i.e., number of lineup members).
Positive predictive value (PPV)	The proportion of suspect identifications in which the suspect's face is a match to the perpetrator's face. PPV is calculated by dividing correct identifications by the sum of correct identifications and innocent suspect identifications.

Introduction

Closed-circuit television (CCTV) cameras can provide valuable evidence in criminal investigations. In addition to showing what happened at the crime scene, CCTV footage can also provide information about who committed the crime. To identify someone from video evidence, an image of a suspect can be compared to the capture of the perpetrator (Moreton, 2021). This type of identification task is an example of face matching, which involves deciding whether two faces are the same person. Face matching is a perceptual task and can be done without any prior memory of the person. This means that investigators can use video evidence to identify perpetrators while circumventing the many pitfalls of relying on eyewitnesses with imperfect memories. Nonetheless, research indicates that matching a face to a video can be challenging and error-prone (Bruce et al., 1999; Bruce et al., 2001; Henderson et al., 2001; Kemp et al., 1997; Megreya & Burton, 2008; Susa et al., 2019; White et al., 2014).

Across numerous jurisdictions, including Canada, the United Kingdom (UK), and the United States (US), the current procedure for identification from video evidence requires only 1-to-1 comparison between a suspect and the capture of the perpetrator (Facial Identification Scientific Working Group, 2022). In the context of testing eyewitnesses, presenting a suspect alone for identification is known as a showup. Due to the suggestiveness of presenting a lone suspect for eyewitness identification, the danger of mistaken identification at showups has been widely recognized by the courts. The preferred procedure for testing eyewitness identification is to present the suspect in a lineup with other individuals, called fillers, who are known to be innocent (Steblay et al., 2019; Wells & Turtle, 1986; Wells et al., 2020). In doing so, lineups offer procedural fairness by concealing the identity of the suspect and consequently reduce suggestiveness (Dysart & Lindsay, 2007; Kassin et al., 2001). Furthermore, according to differential filler-siphoning theory, lineups result in better outcomes than showups because fillers draw more identifications away from innocent suspects than they do from guilty ones (Smith et al., 2017). Therefore, by protecting innocent suspects from mistaken identification, lineups improve the accuracy of suspect identifications.

As technology advances and CCTV becomes more widespread, the use of video evidence to aid criminal investigations will increase; therefore, a scientifically-informed procedure with better safeguards for innocent suspects must be developed for video evidence identification. Drawing from research on eyewitness identification, the lineup method provides a promising solution to the risk of misidentification from video evidence (Steblay et al., 2019; Wells & Turtle, 1986; Wells et al., 2020). The purpose of this thesis is to test whether face matching lineups improve identification outcomes for video evidence over the current practice of 1-to-1 comparison.

Face Matching

When someone is caught committing a crime on camera, police use face matching to compare potential suspects to the person in the video. If police believe the suspect is a match to the perpetrator in the capture (i.e., they are the same person), then police may conclude that the suspect is guilty, and this could result in conviction. However, face matching is error-prone, and participants in experimental research have been found mistaking two different people as the same person at a rate of up to 30% (Bruce et al., 1999; Bruce et al., 2001; Henderson et al., 2001; Kemp et al., 1997; Megreya & Burton, 2008; Susa et al., 2019). Even individuals who have professional experience, such as police officers (With & Carbon, 2017), or those with superior face recognition abilities, such as super-recognizers (Bate et al., 2018; Bobak et al., 2016; Phillips et al., 2018), are prone to making errors in face matching. Most efforts to improve face matching through training have also been unsuccessful (Towler et al., 2019).

In recent years, there have been advances in facial recognition technology, and some police departments use it as an investigative tool. When a crime is caught on camera, a capture from the footage, also known as a probe image, may be submitted by the police to facial recognition technology (Office of the Privacy Commissioner of Canada, 2022). Facial recognition technology then uses algorithms to compare the probe image to a reference database of images, which can include previous mugshots, driver's license photographs, and even social media photographs (Finklea et al., 2023). The result is a rank-ordered list of match candidates, which the police or other human analysts review to decide if any are similar enough to the probe image to investigate further (Finklea et al., 2023).

Despite evidence of some newer algorithms surpassing the accuracy of human observers (Phillips et al., 2018), facial recognition technology is not always accurate. Facial recognition technology is known to perform particularly poorly with ethnic minorities, females, and both young and elderly faces (Lynch, 2020). Therefore, police are required to verify the output produced by facial recognition technology. Using face matching, police must decide whether any of the match candidates are similar enough to the perpetrator in the capture to investigate further (Finklea et al., 2023). However, recent research has shown that humans often fail to correct mistaken match outputs made by facial recognition technology (Carragher & Hancock, 2023). Therefore, even with the combined effort of humans and technology, misidentifications from video evidence can and do occur.

In 2023, Porcha Woodruff became the sixth of seven confirmed people in the US to be wrongfully arrested due to misidentification from video evidence (Sanford, 2024). Woodruff was wrongfully arrested for a carjacking and robbery in Detroit after police determined that her mugshot, which appeared on a candidate list generated by facial recognition technology, was a match to the perpetrator in the capture. Woodruff was eight months pregnant at the time of arrest, and nowhere on the police report did it mention that the perpetrator caught on camera was pregnant (Bhuyian, 2023). The mugshot that appeared on the candidate list was also taken eight years prior. Images used for comparison, such as mugshots or driver's licence photographs, are often taken days, months, or years before the crime is captured, and previous research has shown that face matching performance drastically drops when photographs are taken several months apart (Davis & Valentine, 2009; Megreya et al., 2013; Sandford & Ritchie, 2021). If the crime occurs long after a suspect's image was taken, as in Woodruff's case, there is a higher likelihood of changes in appearance. Consequently, it can be more challenging for investigators to correctly identify the perpetrator.

Identifying someone from video evidence is also made more challenging by the often less-than-ideal conditions of CCTV footage. In the UK, over 80% of the footage obtained from CCTV cameras are of poor-quality for identification (Lee et al., 2009). To cover broad areas, CCTV cameras are placed high up and use wide-angle lenses. Consequently, CCTV footage may show people from odd angles, in low contrast, or in harsh lighting. Images can also look blurry when zooming in on a specific person (Rogers, 2019). It is difficult to see facial features clearly with poor-quality footage.

Therefore, considerable confusion between similar-looking individuals can result, which makes establishing a perpetrator's identity more challenging for both humans and facial recognition technology (Bruce et al., 1999; Henderson et al., 2001; O'Toole et al., 2012).

Even if someone is caught committing a crime on high-quality CCTV footage, it is possible that they would have concealed their face to avoid identification (Gill & Loveday, 2003). Masked images frequently prevent facial recognition technology from processing faces, leaving the identification entirely to police (Ritchie et al., 2024). Unsurprisingly, research has shown that disguises impair facial recognition (Hockley et al.,1999; Terry, 1993; Shapiro & Penrod, 1986). The concealment of external features, such as hairline, is particularly harmful for the recognition of unfamiliar faces (Ellis et al., 1979; Henderson et al., 2001; Young et al., 1985). Even everyday accessories, such as sunglasses or medical masks, can impair face matching performance (Carragher & Hancock, 2020; Dhamecha et al., 2014; Estudillo et al., 2021; Graham & Ritchie, 2019; Kramer & Ritchie, 2016; Ritchie et al., 2024).

The current best-practice guidelines set forth by the Facial Identification Scientific Working Group (FISWG) do not entirely mitigate the risk of misidentification. FISWG was created by the Federal Bureau of Investigation in 2009 with the purpose of developing standards and guidelines for image-based comparisons of human faces. Today, FISWG includes over 50 law enforcement agencies, intelligence agencies, and private industries worldwide (Facial Identification Scientific Working Group, n.d.). FISWG recommends the use of morphological analysis for image-based comparisons of human faces. Morphological analysis involves comparing the shape and form of individual features, such as the nose, and even finer grain details like wrinkles (Moreton, 2021). However, there is limited research validating the use of morphological analysis for video evidence identification (Moreton, 2021). What research has been conducted has shown considerable within-group variability for forensic face examiners who use this technique, with some performing below the average score of untrained individuals (Phillips et al., 2018).

To structure the morphological analysis for 1-to-1 face matching, FISWG (2023) recommends using the Analysis, Comparison, Evaluation, and Verification (ACE-V) method. In the ACE-V method, examiners must first determine whether the image is of sufficient detail and quality for examination. If quality is acceptable, examiners then conduct a visual side-by-side comparison using morphological analysis, taking note of any similarities or dissimilarities, and annotating the images to draw attention to noteworthy features. During the evaluation stage, examiners must consider the value of the dissimilarities and similarities between the two faces to form an opinion on whether they match. In the verification stage, a second independent examiner repeats the analysis, comparison, and evaluation steps. Forensic confirmation bias can occur when the course of a criminal case is influenced by an individual's pre-existing beliefs, expectations, motives, or contextual information (Kassin et al., 2013). To protect against forensic confirmation bias, FISWG recommend the verification stage be blind (without knowledge of the other examiner's opinion).

Although blind testing can minimize the biasing effects of previous conclusions and contextual information related to the case, it cannot eliminate the possible assumption that the suspect is likely to be the perpetrator (Kassin et al., 2013). The current practice for many forensic sciences, including face matching, involves the comparison of a suspect sample (i.e., suspect's face) to the crime scene sample (i.e., the captured face) to determine if they match. This type of 1-to-1 comparison is analogous to presenting the suspect alone to an eyewitness for identification (i.e., a showup; Kassin et al., 2013). Researchers have argued that presenting a single suspect for identification is suggestive because it indicates who the suspect is (Wells et al., 2013). Therefore, even if police and examiners are shielded from contextual information about the case, the suggestiveness of a 1-to-1 comparison could bias them towards making a match decision (Quigley-McBride & Wells, 2018).

Furthermore, research on eyewitness identification has revealed that showups can result in more false alarms when the innocent suspect and perpetrator resemble one another (Steblay et al., 2003). This is particularly applicable to video evidence identification, which is likely to lead to suspicion of either the perpetrator or an innocent suspect of high similarity to the perpetrator (Wells & Penrod, 2011). Police typically start with CCTV footage and investigate suspects because they resemble the perpetrator in the capture. This is especially true if facial recognition technology is used as an investigative tool. Facial recognition technology is programmed to return a candidate list of individuals who exceed a set threshold of similarity to the person in the probe image (Finklea et al., 2023). Therefore, if an innocent person becomes a suspect due to appearing on a candidate list, they will be highly similar to the true perpetrator, putting them at greater risk of misidentification.

The Filler-Control Method

To minimize bias in forensic sciences, experts have recommended using evidence lineups when feasible (Kassin et al., 2013; Saks et al., 2003; Wells et al., 2013). Presenting a suspect in a lineup alongside a group of known innocents is referred to as the filler-control method (Wells et al., 2013). As demonstrated in the eyewitness identification literature, the filler-control method can help to minimize bias in the identification procedure by concealing who the suspect is (Quigley-McBride & Wells, 2018). In addition, because fillers are known innocents, the identification of a filler does not have the same forensic implications as the identification of a suspect. Instead, filler identifications represent known errors and do not pose a risk of wrongful conviction. Presuming that the lineup is fair and that the fillers are plausible alternatives, the filler control method offers better protection to innocent suspects than showups because, by chance, a mistaken identification of one of the fillers is the more likely outcome (Wells & Turtle, 1986). Consequently, the use of lineups with fillers has become standard practice for eyewitness identification across many jurisdictions (Steblay et al., 2019; Wells & Turtle, 1986; Wells et al., 2020). Table 1 provides definitions of key terms related to the filler-control method as well as other key concepts discussed throughout this paper.

Table 1 Definition of key terms

Term	Definition
Fillers	Lineup members who are known to be innocent.
Filler-control method	Presenting a suspect alongside a group of known-innocent fillers.
Nominal size correction	A method for estimating innocent suspect identifications. Involves dividing the overall false alarm rate in target-absent lineups by nominal size (i.e., number of lineup members).
Effective size correction	A method for estimating innocent suspect identifications. Involves dividing the overall false alarm rate in target-absent lineups by effective size (i.e., number of plausible lineup members). A lineup member is plausible if they attract a certain number of identifications.
Positive predictive value (PPV)	The proportion of suspect identifications in which the suspect's face is a match to the perpetrator's face. PPV is calculated by dividing correct identifications by the sum of correct identifications and innocent suspect identifications.
Negative predictive value (NPV)	The proportion of rejections that are correct. NPV is calculated by dividing correct rejections by the sum of correct rejections and incorrect rejections.

Note: Definitions are adapted from Fitzgerald et al., (2023), Smith et al., (2017), Wells et al., (2013), and Wixted & Wells (2017).

Diagnostic feature detection theory and differential filler-siphoning theory provide two explanations for why lineups result in better accuracy than showups. According to diagnostic feature detection theory, simultaneous lineups, which present the lineup members all at once, are better than showups because they improve the witness's ability to discriminate between innocent and guilty suspects. This theory suggests that some facial features are diagnostic of guilt because they differ between innocent and guilty suspects, while others are not diagnostic because they are shared by all lineup members (Wixted & Mickes, 2014). Presuming that the fillers have been selected to match the eyewitness description of the perpetrator, the nondiagnostic features would be those that correspond to that description. This is because the suspect and the fillers would share the described features. Using a simultaneous lineup allows witnesses to evaluate and discount features that are shared amongst all lineup members and focus on features that are more diagnostic of guilt, which is theorized to improve discriminability (Wixted & Mickes, 2014). Showups do not enable this comparative process; therefore, witnesses may depend too heavily on nondiagnostic features, reducing their ability to discriminate between guilty and innocent suspects.

However, Smith and colleagues (2017) have shown that fillers do not actually need to improve discriminability between innocent and guilty suspects to result in better identification outcomes over showups. Instead, they note that fillers attract false alarms (i.e., mistaken identifications) and prevent witnesses from identifying innocent suspects through a process known as filler-siphoning (Wells et al., 2015). Smith and colleagues (2017) argue that filler siphoning is differential, such that fillers siphon more positive choices away from innocent suspects than they do from guilty ones. This is because fillers will be as similar to the witness's memory of the perpetrator as the innocent suspect, if lineups are fair. Therefore, fillers effectively compete for choices when the suspect is innocent. Conversely, the perpetrator will typically be a better match to memory than the fillers. As a result, fillers tend not to compete as effectively for choices when the suspect is guilty (Smith et al., 2017). Consistent with this theoretical account, research has demonstrated that lineups yield significantly lower rates of innocent suspect identifications than showups, and in some cases similar rates of correct identifications (Dysart & Lindsay, 2007; Gronlund et al., 2012; Steblay et al., 2003; Valentine et al., 2012). However, because of filler-siphoning, lineups can result in fewer correct identifications than showups (Beal et al., 1995; Gonzales et al., 1993; Smith et al., 2023).

Quigley-McBride and Wells (2018) applied the filler-control method to a forensic fingerprint analysis task and found that it led to a reduction in false alarms and a slight reduction in hits when compared to the standard single suspect method. Consistent with differential filler-siphoning theory, the reduction in false alarms was not a product of increased correct rejections, but rather filler-siphoning. These findings demonstrate that applying the filler-control method to forensic perceptual tasks can yield the same benefits as the ones shown for eyewitness identification. Therefore, using the filler-control method for video evidence identification, which is also a forensic perceptual task, may result in better protection for innocent suspects.

Lineups can also result in lower expected costs than showups (Yang et al., 2019). The cost of an identification decision refers to its discrepancy to the goals of the investigation. The goal of an investigation is almost always to apprehend the perpetrator;

therefore, if the perpetrator is correctly identified, there is zero cost. For all identification decisions that fail to meet this goal, there is an added cost of failing to incriminate the perpetrator (Steblay et al., 2011; Wells et al., 2012). Wrongfully incriminating an innocent suspect can be particularly costly, depending on societal values. If societal beliefs reflect Blackstone's view that it is "better that ten guilty persons escape, than that one innocent suffer," (1769, p. 353), then the incrimination of an innocent suspect may be viewed as more costly.

An ideal identification procedure would reduce mistaken identifications without also reducing correct identifications. However, certain reforms can result in a trade-off between mistaken and correct identifications (Clark, 2012; Juncu & Fitzgerald, 2021; Wetmore et al., 2017; Wooten et al., 2020). The impact of this trade-off is dependent on the prior probability of guilt, which refers to the probability that the suspect is guilty before conducting the identification procedure. In the real world, the prior probability of guilt is challenging to assess, and it can vary between jurisdictions, depending on existing policies (Wells & Olson, 2003). The cost of an identification procedure that results in reduced mistaken and correct identifications will be greater when the prior probability of guilt is high. In comparison to showups, lineups have been shown to decrease mistaken identifications. However, lineups do not always have a favourable effect on correct identifications. Although some research has shown that lineups and showups yield similar rates of correct identifications (Dysart & Lindsay, 2007; Gronlund et al., 2012; Steblay et al., 2003; Valentine et al., 2012), there is evidence that lineups can result in fewer correct identifications than showups due to filler-siphoning (Beal et al., 1995; Gonzales et al., 1993; Smith et al., 2023). Having said that, if lineups decrease mistaken identifications without impacting correct ones, then the use of lineups rather than showups will not result in a trade-off and the expected cost of a lineup will be smaller than that of a showup, regardless of prior probability of guilt (Yang et al., 2019).

While lineups may not improve discriminability, an additional identification measure, known as the rule-out lineup procedure, has been previously found to improve discrimination between guilty and innocent suspects over both showups and traditional lineups (Ayala et al., 2022; Smith & Ayala, 2021). Consistent with a traditional lineup, the rule-out procedure begins by asking the witness to make an identification decision and to express their confidence in that decision. After confidence for the decision is recorded, the witness is asked to indicate how confident they are that any non-identified lineup members are, indeed, not the perpetrator (Smith & Ayala, 2021). Essentially, the witness is asked to rate their confidence in the innocence of each non-identified lineup member. This means that even when a filler is selected or the lineup is rejected, a confidence measure that indicates how well the suspect matches the witness' memory is obtained. Consequently, in comparison to lineups alone, the rule-out procedure maximizes the potential for eyewitness memory to both inculpate the guilty and exculpate the innocent (Ayala et al., 2022; Smith & Ayala, 2021).

Despite lineups being widely recognized as best practice for eyewitness identification and research showing benefits of lineups for forensic perceptual tasks, FISWG do not recommend lineups for identification from video evidence. This could be because there is no research supporting the use of the lineups for face matching tasks. In both the face matching literature and in video evidence identification cases, performance is typically assessed with a 1-to-1 task. This involves deciding whether two faces are the same person. An alternative method that has been used in some experimental studies is the one-to-many task (Megreya & Burton, 2007; Megreya & Burton, 2008; Megreya et al., 2013). In the one-to-many task, the participant is asked to compare a single target face to a group of potential matches, so it is procedurally the same as a face matching lineup. However, despite this previous research with the one-to-many task, it remains unclear whether face matching lineups can improve outcomes over the 1-to-1 procedure. This is because rather than testing the potential for lineup fillers to protect an innocent suspect, the one-to-many task has been traditionally used to learn about the cognition involved in face matching. To our knowledge, 1-to-1 and lineup procedures for face matching have not been directly compared to each other in a single experiment. Furthermore, the results of previous studies that employed the one-to-many task have not been interpreted from the perspective of the filler-control method.

The filler-control method explains why a face matching lineup is likely to have applied benefits for video evidence identification. The way to measure the benefits of this approach is to compute the positive predictive value (PPV) of suspect identifications. PPV is a measure that is of great interest to the criminal justice system because it is used to assess the reliability of an identification technique (Mickes, 2015; Smith & Neal, 2021). In the context of video evidence identification, PPV refers to the proportion of suspect identifications in which the suspect's face actually matches the perpetrator's face. PPV is calculated by dividing correct identifications by the sum of correct identifications and innocent suspect identifications (Smith et al., 2017). Research on eyewitness identification has shown that, because of differential filler-siphoning, lineups have better PPV than showups (Smith et al., 2017). When differential filler-siphoning occurs in lineups, fillers detract more choices away from innocent suspects than guilty ones, reducing innocent suspect identifications, without greatly impacting correct identifications, which ultimately leads to greater PPV.

To realize the benefits of lineups on PPV, it is critical to distinguish between mistaken identifications of fillers and mistaken identifications of an innocent suspect. I reanalyzed the false alarm rates of three face matching studies that used a 1-to-10 task (Megreya & Burton, 2007; Megreya & Burton, 2008; Megreya et al., 2013), using a nominal size correction. The nominal size correction is used in eyewitness identification research to estimate innocent suspect identifications by dividing the overall false alarm rate in target-absent lineups by the number of lineup members (Wixted & Wells, 2017). Given that the task included 10 options, I divided the false alarm rates from target-absent trials by 10. This caused the reported false alarm rates to decline from .26, .21, and .33 to .03, .04, and .03, respectively. The corresponding hit rates were .72, .71, and .78. Before applying the nominal size correction, the respective PPVs were estimated to be .73, .77, and .70. After applying the nominal size correction, PPV increased to .97, .95, and .96. However, by applying the nominal size correction, we are assuming that lineups are fair, and all 10 options are equally as likely to be selected as a match (Fitzgerald et al., 2023). Therefore, if fair lineups are assumed, employing the filler-control method with face matching lineups could result in higher PPV than 1-to-1 comparisons.

Current Study

Research on eyewitness identification suggests that the current procedure for video evidence identification does little to protect innocent suspects from misidentification. For this reason, I tested the use of lineups for face matching with video evidence. To test the filler-control method in the face matching context, participants completed a video identification task using two procedures: the 1-to-1 procedure, where an image of the suspect was compared to a video still of the perpetrator, and the face matching lineup, where a photo lineup comprised of one suspect and five fillers was compared to the video still of the perpetrator. I also assessed performance on the rule-out procedure. By testing the filler-control method in the face matching context, this thesis aims to contribute to the development of a procedure for video evidence identification that prevents wrongful conviction.

Hypothesis

In my pre-registered hypothesis (https://aspredicted.org/DRV_LYC) I predicted that if the nominal size correction was used to estimate the risk to innocent suspects, face matching lineups would result in better PPV than 1-to-1 comparisons.

Method

Participants

Undergraduate students (*N* = 591) were recruited through a Psychology Department recruitment system to complete an online video identification task. They received partial course credit for their participation. For a paired sample *t*-test, the power analysis indicated that a minimum sample size of 327 would be necessary to achieve 95% power to detect a small (*d* = 0.20) effect size.

The final sample included 541 participants (71% identified as female, 28% as male, 1% as non-binary). In terms of age, 98.1% were 18-24, 1.7% were 25-34, and 0.2% were 45-54. For ethnicity, 34.6% identified as White, 26.4% as South Asian, 26.2% as East Asian, 5.1% as Middle Eastern, 4.2% as Mixed, 1.9% as Hispanic, 0.7% as Black, 0.5% as Indigenous, and 0.5% as African. Fifty participants were excluded due to missing data, four of which exited the Qualtrics survey immediately after consenting.

Due to programming errors, confidence for target and filler identifications is missing for 127 target-present lineup trials. Additionally, there are 15 target-absent and 19 target-present lineup trials where participants skipped the rule-out procedure that followed. Therefore, data for 34 rule-out trials are missing.

Design

This study employed a 2 (target: present or absent) × 2 (procedure: 1-to-1 or lineup) within-subjects design. Over 28 trials, divided into two blocks, participants judged whether a perpetrator in a crime video image matched a suspect mugshot image. The use of 1-to-1 or lineup was manipulated between blocks, and participants completed seven target-present and seven target-absent trials for both procedures. The dependent variables of interest were correct identifications on target-present trials, innocent suspect identifications on target-absent trials, positive predictive value, negative predictive value, area under the curve (AUC), and confidence.

Materials

Crime Video Images and Perpetrator Mugshots

Figure 1 shows an example of a target-present crime video image and a perpetrator mugshot. Figure 2 shows an example of a target-absent crime video image and an innocent suspect mugshot. Crime video images and perpetrator mugshots were selected from the Eyewitness Lineup Identity (ELI) stimulus database, which includes multiple images and videos of approximately 200 individuals (Fitzgerald et al., 2023). All individuals included in the ELI stimulus database volunteered and consented to having their photographs taken. For this study, 53.5% of the individuals selected from the ELI stimulus database were female and 46.5% were male. There is a crime video clip and a mugshot image for each individual in the ELI stimulus database.

For each trial, participants saw one crime video image. Some CCTV cameras can zoom, and police likely utilize this function to get a better view of the perpetrator's face (Surrette, 2004). Even if CCTV cameras are not equipped with zooming capability, police likely zoom-in on the perpetrator once they have obtained a capture from the footage. Therefore, crime video images were created by taking a screenshot from the video clips and zooming in on the perpetrator, so that they could only been seen from the waist up. This allowed participants to have a better view of the perpetrator's face. To cover broad areas, CCTV cameras use wide-angle lenses, which can result in a blurrier image when zooming in on a specific person (Rogers, 2019). Research has shown that face matching is more difficult when the quality of CCTV footage is poor (Henderson et al., 2001). Therefore, crime video images were blurred by 5 percent using the photo-editing platform, Fotor.

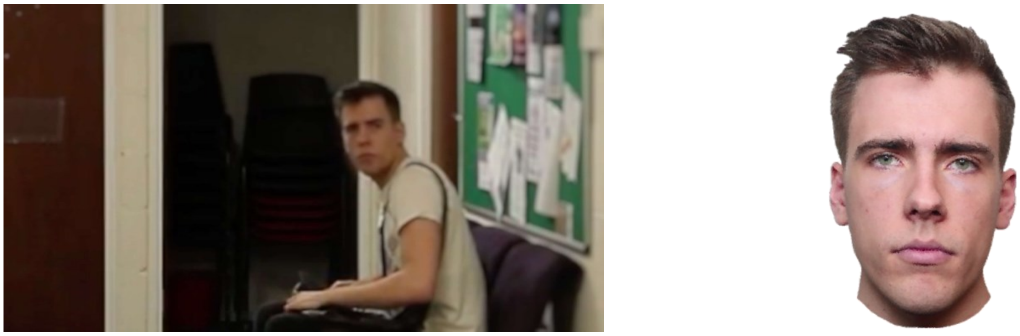

Figure 1 Matching crime video and mugshot image
Note. The image on the left is an example of a crime video image. The image on the right is an example of a suspect mugshot. The crime video image and the suspect mugshot show the same person.

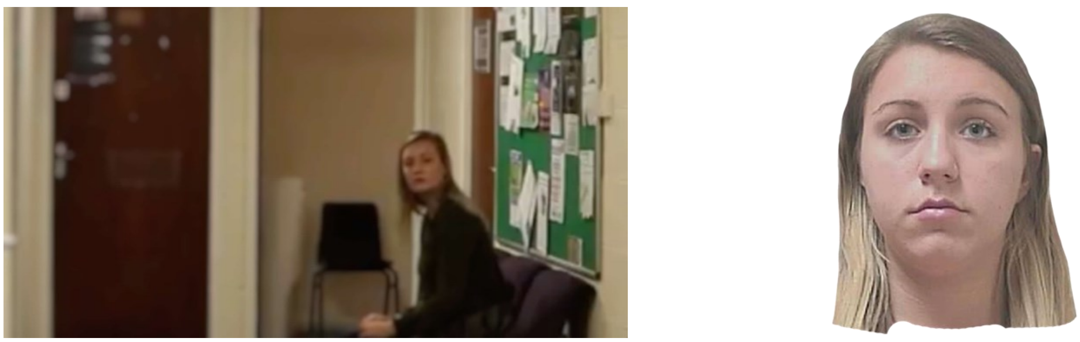

Figure 2 Mismatching crime video and mugshot image
Note. The image on the left is an example of a crime video image. The image on the right is an example of a suspect mugshot. The crime video image and the suspect mugshot show different people.

Lineup Condition

For each trial in the lineup condition participants saw seven images presented simultaneously: (1) a still image from the crime video footage, and (2) a six-person lineup presented in a 2 × 3 array. In addition to the seven images, participants also saw a button that said, “There is no mugshot image that matches the perpetrator caught on camera”.

I used 28 lineups from a previous study that were constructed to be fair (Smith et al., under review). Smith and colleagues also created biased lineups, but none of the biased lineups were tested in this thesis. Perpetrator mugshots were selected from the ELI stimulus database. Filler photographs were selected from a Florida mugshot database. To construct lineups, a team of research assistants were instructed to find six fillers that both matched a description of the perpetrator and generally resembled the perpetrator. The quality of the perpetrator mugshot was degraded to match that of the filler photographs. All lineup photographs were edited to remove backgrounds and clothing (e.g., shirt collars).

Target-present lineups included the guilty suspect plus five fillers. Target-absent lineups comprised six fillers, none of which were designated as an innocent suspect. This is because I did not want innocent suspect identifications to be influenced by any subjective criteria used to select an innocent suspect. Furthermore, the designation of an innocent suspect would have been arbitrary in this experiment because I counterbalanced the images such that all six lineup members were rotated (between subjects) to appear in the 1-to-1 comparison. To counterbalance each condition, I created target-present and target-absent 1-to-1 trials and lineup trials for each crime video image. To equate target-absent trials, each filler in the lineup was rotated into the innocent suspect position of the corresponding 1-to-1 trial, which showed the same crime video image. A Qualtrics randomizer was used to ensure each filler was rotated into the innocent suspect position of the 1-to-1 trial an equal number of times. Rotating fillers from the lineup into the 1-to-1 trial was not necessary for target-present trials, as the suspect was always the perpetrator. Table 2 shows how I counterbalanced 1-to-1 and lineup trials by creating four different versions of the study from four different stimulus sets (A, B, C, and D). The number of innocent suspect identifications was instead estimated using the nominal size correction and effective size correction (Fitzgerald et al., 2023). The effective size correction is explained in Data Analysis.

1-to-1 Condition

Each 1-to-1 trial comprised two simultaneously presented images: (1) a still image from the crime video footage, and (2) a mugshot image of the suspect. In addition to the two images, participants also saw a button that said, “The mugshot image does not match the perpetrator caught on camera”. For half of the trials in the 1-to-1 condition, the person depicted in the suspect mugshot was the same as the person in the crime video image (target-present condition). For the other half of the trials, the suspect mugshot showed a different person from the crime video image (target-absent condition). The 1-to-1 trials were constructed using the same images from the lineup trials. Therefore, each target-present trial showed a perpetrator mugshot from the ELI stimulus database and each target-absent trial showed a filler mugshot from the Florida mugshot database.

Table 2 Counterbalancing of 1-to-1 and lineup trials

Version	Target-present lineups	Target-absent lineups	Target-present 1-to-1s	Target-absent 1-to-1s
1	Stimulus set A	Stimulus set B	Stimulus set C	Stimulus set D
2	Stimulus set B	Stimulus set C	Stimulus set D	Stimulus set A
3	Stimulus set C	Stimulus set D	Stimulus set A	Stimulus set B
4	Stimulus set D	Stimulus set A	Stimulus set B	Stimulus set C

Practice Trials

Participants were required to complete one practice trial in each condition before being tested. Like real trials, the 1-to-1 practice trial was constructed using images from the ELI stimulus database, while the lineup trial was constructed using images from the ELI stimulus database and Florida mugshot database. For both conditions the practice trial was target-present and showed the same crime video image and perpetrator mugshot.

Procedure

The study was completed online using Qualtrics. Participants were required to use a PC or laptop to complete the study. If participants attempted to use a mobile device, they were redirected to the end of the survey and instructed to try again using a PC or laptop. This was to ensure image size was consistent between participants.

Participants were instructed that they were to complete two shifts as a criminal investigator where their job was to identify perpetrators who have been caught committing crimes on camera. Participants were told (a) the task was to decide whether or not a suspect is the same person as the perpetrator caught on camera; (b) suspects may or may not be guilty (c) guilty suspects match the perpetrator, and innocent suspects do not match the perpetrator; and (d) be as accurate as possible.

Participants completed two blocks, one for the 1-to-1 condition and one for the lineup condition. Each block consisted of 14 trials (i.e., each participant completed 28 trials total), and half of the trials contained guilty suspects (i.e., target-present) and the other half contained innocent suspects (i.e., target-absent). Each trial could be viewed for as long as needed. The order of the blocks and trials were randomized between participants. Before starting each block, participants were provided with instructions on how to complete the identification procedure and shown an example. Participants were required to complete one practice trial following the instructions. Participants were also asked to indicate their understanding of the instructions by selecting a specific multiple-choice option (understanding check question). Participants were able to skip the understanding check question. If a participant skipped this question and performed poorly throughout the study, this may indicate that the participant was not paying attention. All participants correctly answered the understanding check questions.

For each trial in the 1-to-1 condition, a single mugshot image appeared next to the crime video image of the perpetrator. Participants were asked "Does the mugshot image show the same person as the perpetrator in the crime video image?". If participants believed that the mugshot image matched the perpetrator in the crime video image, they were instructed to click on the mugshot image. If they believed that the mugshot image did not match the perpetrator, they were instructed to select the button that said, "The mugshot image does not match the perpetrator caught on camera". Participants were then asked to rate their confidence in the decision they made on a scale of 0-100. Participants were able to view the images and their response when rating their confidence.

For each trial in the lineup condition, a lineup of six mugshots appeared next to the crime video image of the perpetrator. Participants were instructed that there was only one suspect in each lineup, and they would not know who the suspect was. For each trial in this condition, participants were asked "Do any of the mugshot images show the same person as the perpetrator in the crime video image?". Participants were told “If you believe a mugshot image matches the perpetrator (i.e., they are the same person), then you must click on that mugshot image.” If participants believed none of the mugshot images matched the perpetrator, they were instructed to select the button that said, "There is no mugshot image that matches the perpetrator caught on camera". Participants were then asked to rate their confidence in the decision they made on a scale of 0-100. Again, participants were able to view the images and their response when rating their confidence.

After making this initial decision, participants were then shown the same crime video image and lineup for a second time to complete the rule-out procedure. During this second viewing, participants indicated how confident they were, on a scale of 0-100, that each non-selected lineup member is not the perpetrator in the crime video image. There was a textbox below each lineup member that was not initially selected as a match where participants were required to input their confidence that the person was not the perpetrator. Using Qualtrics' display logic feature, the rule-out procedure trials were programmed so that textboxes only appeared below lineup members who were not initially selected as a match. If a lineup member was initially selected as a match there was a label above their mugshot that read “This is the mugshot you selected”. If participants initially decided that no mugshot image matched, they had to provide a confidence rating for all lineup members during the second viewing.

Once both shifts were complete, participants were asked to answer demographic questions regarding their age, ethnicity, sex, and gender identity. Participants were not required to answer the demographic questions if they did not want to. Participants were then debriefed and thanked for their participation.

Data Analysis

Accuracy

Correct and incorrect response rates were calculated for each condition by dividing total responses by the number of trials (7). In both conditions, correct responses include correct identification and correct rejection. A correct identification refers to correctly selecting the culprit as a match on target-present trials. A correct rejection refers to correctly responding that the innocent suspect is not a match on target-absent trials. In the 1-to-1 condition there are only two possible incorrect responses: incorrect rejection and false alarm. An incorrect rejection refers to incorrectly responding that the guilty suspect is not a match on target-present trials. A false alarm refers to incorrectly selecting a nonmatching individual as a match on target-absent trials. On target-absent 1-to-1 trials, the only nonmatching individual shown was the innocent suspect; therefore, all false alarms were innocent suspect identifications. In the lineup condition, because we did not designate an innocent suspect, there were also only two possible incorrect responses participants could make: incorrect rejection and false alarm. Overall false alarms were equivalent to the total number of fillers identified on target-absent lineups.

To estimate the number of innocent suspect identifications I used two methods: nominal size correction and effective size correction. The nominal size correction involves dividing filler identifications, or overall false alarms, by lineup size (6). One caveat of the nominal size correction is its assumption that lineups are perfectly fair and that the chances of mistaken identification are the same for each lineup member (Fitzgerald et al., 2023).

I also estimated the number of innocent suspect identifications using the effective size correction, which is sensitive to the distribution of lineup choices (Smith et al., 2021; Quigley-McBride & Wells, 2021). The effective size correction involves dividing overall false alarms by the effective size, which is the number of plausible lineup members (Malpass, 1981). A lineup member is considered plausible if they attract a certain number of identifications. Therefore, rather than assuming the lineup is fair, this method assumes the innocent suspect is one of the plausible lineup options (Fitzgerald et al., 2023). To measure effective size, I used Tredoux's (1998) formula. If lineups are perfectly fair and there is an even distribution of identifications across lineup members, then the nominal size and effective size correction will result in the same number of innocent suspect identifications (Fitzgerald et al., 2023). Conversely, if lineups are less than perfectly fair, and identifications are not distributed evenly, then the number of innocent suspect identifications estimated by the effective size correction will be greater than the number estimated by the nominal size correction. This is because, under imperfect conditions, the effective size of a lineup would be less than the nominal size of the lineup.

Positive predictive value

For each condition, I calculated PPV using the following formula: correct identifications/(correct identifications + innocent suspect identifications). For the lineup condition, I calculated three different PPVs: one using the total number of false alarms, one using the innocent suspect identifications estimated by the nominal size correction, and one using innocent suspect identifications estimated by the effective size correction. The purpose of using the total number of false alarms from lineups to calculate PPV was to assess performance in situations where no filler-siphoning would occur, such as when all lineup members are suspects. To test my hypothesis that lineups would result in better PPV than the 1-to-1 procedure, I used a paired samples *t*-test to compare lineup PPV that was calculated using innocent suspect identifications estimated by the nominal size correction to 1-to-1 PPV.

Negative predictive value

For each condition, I also calculated negative predictive value (NPV) using the following formula: correct rejections/(correct rejections + incorrect rejections). NPV indicates the capacity of a procedure to exculpate or rule out a suspect (National Research Council, 2014; Smith et al., 2017). To test for differences in NPV, I used a paired samples *t*-test to compare lineup NPV to 1-to-1 NPV.

Receiver operating characteristic (ROC) analysis

To assess investigator discriminability across a range of decision criteria (Wixted & Mickes, 2012), ROC curves were created for the 1-to-1, lineup and rule-out conditions. The discriminability (*d'*) of an identification procedure refers to how well it can distinguish between a signal (i.e., faces that match the perpetrator) and noise (i.e., faces that do not match the perpetrator).

ROC curves are created by plotting responses at different levels of confidence (Wixted & Mickes, 2012). For simplicity, confidence ratings were categorized as low (0-59), medium (60-89), and high (90-100). The leftmost point represents responses that occur at the highest level of confidence. The next point includes a cumulation of responses occurring at the highest and second to highest levels of confidence, and this process continues until a single point in the rightmost region of the curve reflects all responses collapsed over all levels of confidence (Smith et al., 2017). The procedure that results in the ROC curve that is closest to the uppermost left corner is the procedure that yields the best discriminability (Smith et al., 2020). To quantify this, the area under the ROC curve was measured. The larger the area under the curve, the better discriminability the procedure yields.

1-to-1 Task

The standard ROC curve was designed to be used with binary classification tasks and was therefore suitable for plotting all responses from the 1-to-1 condition (Smith et al., 2020). As such, I plotted the cumulative proportions of suspect identifications from high (90-100) to low (0-59) confidence and cumulative proportions of rejections from low (0-59) to high (90-100) confidence.

Lineup

Standard ROC curves, which only consider suspect identifications and rejections, undermine the informational value of lineups by not including all responses. Investigator discriminability refers to how well investigators can distinguish guilty suspects from innocent ones given the information provided from the identification procedure (Smith et al., 2020). To capture investigator discriminability and the true informational value of a lineup, I plotted a full ROC curve, which considers all possible responses that can be used to inform a suspect's guilt, including filler identifications (Smith et al., 2020).

For the purposes of the ROC analysis, I randomly designated one filler as the innocent suspect for each target-absent lineup. To designate an innocent suspect in target-absent lineups I sorted the datafile by target and then by participant number. I then assigned a label between 1 to 6, which corresponded to the six lineup members for that target. The labels were assigned in numerical order and then repeated (i.e., 1, 2, 3, 4, 5, 6, 1, 2, 3...). For example, the lowest participant number for Target 1 was assigned a label of 1, the second lowest was assigned a label of 2, etc. This resulted in each of the six fillers for a target to be designated as the innocent suspect 22-23 times. This corresponds very closely to the number of times each filler was selected to appear on a target-absent trial for 1-to-1 procedure (i.e., 20-24 times).

To plot the full ROC curve for the face matching lineup condition, I used the a priori ordering described by Ayala and colleagues (2022). The operating points were ordered as follows: suspect identifications from high to low confidence, all filler identifications, rejections from low to high confidence. Filler identifications were collapsed into a single operating point because filler identifications at all levels of confidence are as diagnostic of innocence as low-confidence rejections (Smith & Ayala, 2021).

Rule-out procedure

To create full ROC curves for the rule-out procedure, I used the method created by Smith and colleagues (2022). Following this method, I created a cumulative 6-point scale that combined the confidence that the suspect was the perpetrator (high, medium, low) for participants who initially identified the suspect, and the confidence that the suspect was not the perpetrator, for participants who initially identified a filler or rejected the lineup (low, medium, high). The cumulative distribution of target-present responses was then plotted against the cumulative distribution of target-absent responses on the 6-point scale. This combined 6-point scale ranged from high-confidence that the suspect matched the perpetrator to high confidence that the suspect did not match the perpetrator.

Confidence-accuracy characteristic (CAC) analysis

Lastly, I plotted PPV CAC curves to observe the relationship between confidence and suspect identification accuracy in the lineup and 1-to-1 conditions. PPV CAC curves are graphed by plotting PPV at differing levels of confidence; perfect calibration exists when the expressed confidence reflects the percentage of individuals who are correct (Mickes, 2015). Again, I categorized confidence as low (0-59), medium (60-89), and high (90-100). For the lineup condition, I plotted three different CAC curves: one where I estimated the innocent suspect identifications using the nominal size correction, one where I estimated the innocent suspect identifications using the effective size correction, and one where the overall false alarms was used in place of innocent suspect identifications. The CAC curves were plotted together to illustrate how suspect identification accuracy can be impacted by assumptions made when estimating innocent suspect identifications (Fitzgerald et al., 2023).

Using a similar method, I also plotted NPV CAC curves to observe the relationship between confidence and rejection accuracy in the 1-to-1, lineup, and rule-out conditions. NPV CAC curves are graphed by plotting NPV at differing levels of confidence. Again, I categorized confidence as low (0-59), medium (60-89), and high (90-100).

Results

Across all trials (28 trials), the proportion of correct responses (i.e., correct rejections and correct identifications) for both conditions combined, was .67 (range: .32–1.00), with higher accuracy in the 1-to-1 condition (*M* = .81, *SD* = .11) than in the lineup condition (*M* = .53, *SD* = .16). Rejection, overall false alarm, and suspect identification rates for the target-present and target-absent 1-to-1 and lineup conditions are summarised in Table 3.

Preregistered Hypothesis Test

When I applied the nominal size correction, PPV was significantly higher in the lineup condition (*M* = .85, *SD* = .11) than in the 1-to-1 condition (*M* = .82, *SD* = .13), $t(539) = 4.47$, $p < .001$, $d = 0.24$ [95% CI: 0.14, 0.36], which is consistent with my hypothesis.

Exploratory Analyses

Correct Identifications

A paired *t*-test revealed that the correct identification rate on target-present trials was significantly higher in the 1-to-1 condition (*M* = .83, *SD* = .16) than in the lineup condition (*M* = .65, *SD* = .20), $t(540) = 19.52$, $p < .001$, $d = 0.97$ [95% CI: 0.85, 1.09].

Positive Predictive Value

PPV, assuming all-suspect lineup

When I assumed lineups were comprised only of suspects, all false alarms on target-absent lineups were considered innocent suspect identifications. On target-absent trials, the overall false alarm rate was significantly higher in the lineup condition (*M* = .59, *SD* = .25) than in the 1-to-1 condition (*M* = .20, *SD* = .18), $t(540) = 34.58$, $p < .001$, $d = 1.78$ [95% CI: 1.61, 1.93].

When I used the overall false alarm rate to compute PPV for the lineup condition (*M* = .54, *SD* = .16), it was found to be significantly lower than PPV for the 1-to-1 condition (*M* = .82, *SD* = .13), $t(539) = 37.70$, $p < .001$, $d = 1.98$ [95% CI: 1.80, 2.16].

PPV, assuming innocent suspect is among the plausible lineup options

The average effective size in target-absent lineups was 2.89, with the lowest effective size being 1.42 and the highest being 5.01. Therefore, lineups typically had approximately three plausible options. I estimated innocent suspect identifications by dividing by overall target-absent lineup false alarm rate by the effective size of the lineup (effective size correction), which assumes an innocent suspect would be one of the plausible lineup options. When I applied the effective size correction, the innocent suspect identification rate was significantly higher in the lineup condition (*M* = .23, *SD* = .10) than in the 1-to-1 condition (*M* = .20, *SD* = .18), $t(540) = 3.38$, $p < .001$, $d = 0.17$ [95% CI: 0.07, 0.28].

When I computed PPV using innocent suspect identifications estimated by the effective size correction, PPV in the lineup condition (*M* = .74, *SD* = .16) was significantly lower than PPV in 1-to-1 condition (*M* = .82, *SD* = .13), $t(539) = 12.90$, $p < .001$, $d = 0.70$ [95% CI: 0.58, 0.82].

PPV, assuming perfect lineup fairness

When I assumed perfect lineup fairness and estimated innocent suspect identifications by dividing overall false alarms by nominal size (6; nominal size correction), a paired *t*-test revealed that the innocent suspect identification rate was significantly lower in the lineup condition (*M* = .10, *SD* = .04) than in the 1-to-1 condition (*M* = .20, *SD* = .18), $t(540) = 13.79$, $p < .001$, $d = 0.71$ [95% CI: 0.59, 0.82]. As reported previously (see preregistered hypothesis test), when I computed PPV using innocent suspect identifications estimated by the nominal size correction, PPV was significantly higher in the lineup condition than in the 1-to-1 condition.

Negative Predictive Value

A paired *t*-test revealed that the NPV was significantly higher in the 1-to-1 condition (*M* = .84, *SD* = .13) than in the lineup condition (*M* = .78, *SD* = .23), $t(540) = 9.40$, $p < .001$, $d = 0.51$ [95% CI: 0.40, 0.63].

Table 3 Face matching accuracy rate for 1-to-1 and lineup conditions

Condition	Target	Reject	Suspect ID		
			Uncorrected	Nominal	Effective
1-to-1	Present	.17 (.16)	.83 (.16)	NA	NA
	Absent	.80 (.18)	.20 (.18)	NA	NA
Lineup	Present	.13 (.16)	.65 (.20)	NA	NA
	Absent	.41 (.25)	.59 (.25)	.10 (.04)	.23 (.10)

Note. All values are the average rates. Values in parentheses represent standard deviations. This table shows three different suspect identification rates. Uncorrected represents the uncorrected suspect identification rate. On target absent trials, the uncorrected suspect ID rate refers to the overall false alarm rate. On target-present trials, the uncorrected suspect ID rate refers to hits. Nominal refers to the innocent suspect identification rate that was estimated using the nominal size correction. Effective refers to the innocent suspect identification rate that was estimated using the effective size correction.

Receiver operating characteristic (ROC) curves

To assess investigator discriminability across a range of decision criteria, ROC curves were created by plotting the cumulative responses for each condition at different levels of confidence, (0-59, 60-89, and 90-100). Figure 3 shows the ROC curves for the 1-to-1, the lineup, and the rule-out conditions.

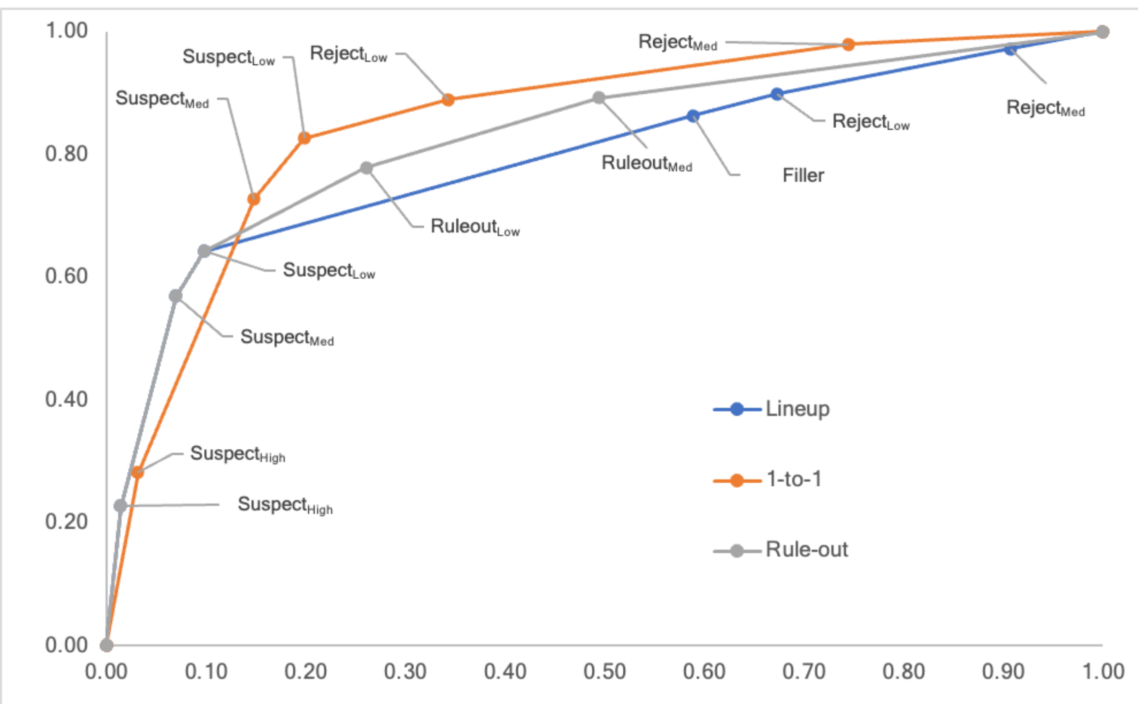

Figure 3 Receiver operating characteristic curves
Note. Innocent suspects were randomly designated for the lineup and rule-out conditions. All points were plotted cumulatively. $\text{Suspect}_{\text{High}}$ refers to suspect identifications made with high confidence (i.e., 90-100), $\text{Suspect}_{\text{Med}}$ refers to suspect identifications made with medium

confidence (i.e., 60-89), and $\text{Suspect}_{\text{Low}}$ refers to suspect identifications made with low confidence (i.e., 0-50). Filler refers to all filler identifications (collapsed over all levels of confidence) in the lineup condition. $\text{Reject}_{\text{Low}}$ refers to rejections made with low confidence and $\text{Reject}_{\text{Med}}$ refers to rejections made with medium confidence. $\text{Ruleout}_{\text{Low}}$ refers to low confidence rule-out ratings that the suspect was not a match to the perpetrator and $\text{Ruleout}_{\text{Med}}$ refers to medium confidence rule-out ratings that the suspect was not a match to the perpetrator.

Area under the curve (AUC)

I computed the AUC for each ROC and compared them using the *roc* test from the pROC package in R. As shown in Figure 3, the 1-to-1 procedure (AUC = .85) was superior to the lineup procedure (AUC = .79) and the rule-out procedure (AUC= .83), $p < .001$. The rule-out procedure was also significantly better than the lineup procedure, $p < .001$.

Confidence-accuracy characteristic (CAC) curves

PPV CAC curves for the 1-to-1 and lineup condition are depicted in Figure 4 to show suspect identification accuracy at low (0-59), medium (60-89), and high (90-100) confidence (Mickes, 2015). For these curves, accuracy was operationalized as the PPV of suspect identifications. For the lineup condition, I plotted three different CAC curves: one where I assumed perfect lineup fairness and estimated the innocent suspect identifications using the nominal size correction, one where I assumed the suspect was one of the plausible lineup options and estimated the innocent suspect identifications using the effective size correction, and one where all false alarms were assumed to be innocent suspect identifications.

The CAC curves in Figure 4 shows that, for the 1-to-1 condition, accuracy increased with confidence. In the 1-to-1 condition, suspects identified with high confidence were 89.7% accurate. Contrarily, high confidence identifications from lineups were only indicative of high accuracy when innocent suspect identifications were estimated using the nominal size correction. When the nominal size correction was applied, suspects identified with high confidence from lineups were 91.7% accurate. However, when the effective size correction was applied, suspects identified with high confidence from lineups were 85.9% accurate. When no correction was applied, suspects identified with high confidence from lineups were only 70.6% accurate.

NPV CAC curves for the 1-to-1, lineup, and rule-out are depicted in Figure 5 to show rejection accuracy at low (0-59), medium (60-89), and high (90-100) confidence (Mickes, 2015). For these curves, accuracy was operationalized as NPV of rejections. The NPV CAC curves in Figure 5 show that high confidence rejections from 1-to-1 trials were 92.4% accurate. Contrarily, high confidence rejections from lineups were only 77.7% accurate. Additionally, the flat slope of the lineup CAC indicates a poor confidence-accuracy relationship for lineup rejections. However, rule-out ratings were more indicative of accuracy with 86.1% of high confidence rule-out ratings being accurate.

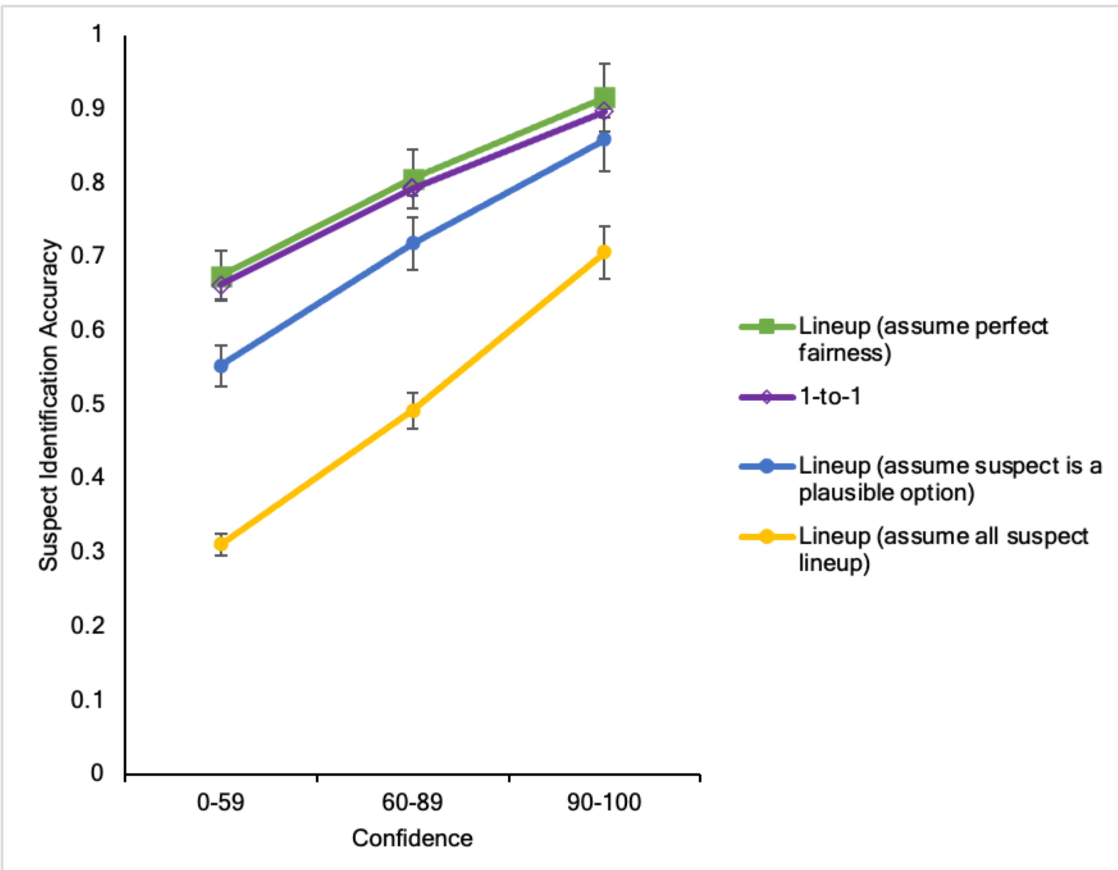

Figure 4 Confidence-accuracy characteristic curves: Positive predictive value

Note. Lineup (assume all suspect lineup) refers to the CAC that was plotted using overall false alarms from target-absent lineups. Lineup (assume suspect is a plausible option) refers to the CAC that was plotted using the innocent suspect identifications that were estimated by dividing overall false alarms by effective size. Lineup (assume perfect fairness) refers to the CAC that was plotted using the innocent suspect identifications we estimated by dividing overall false alarms by nominal size (6).

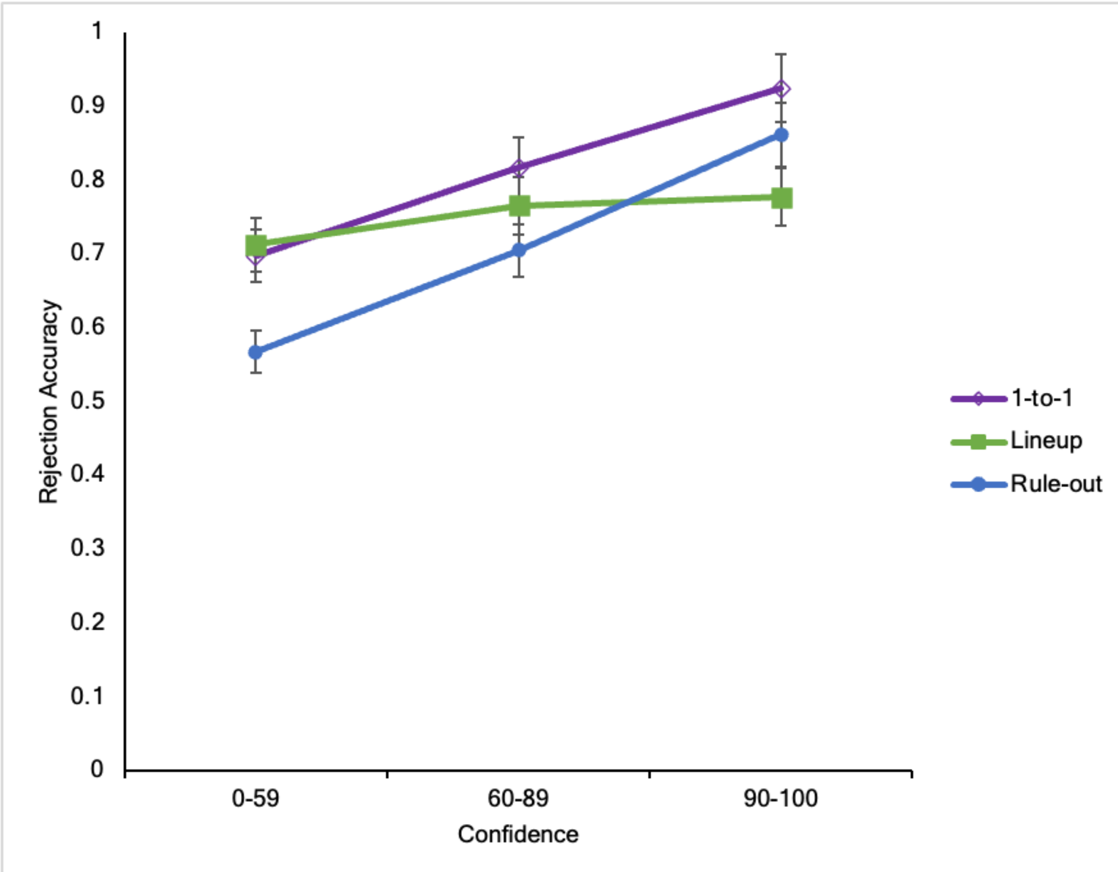

Figure 5 Confidence-accuracy characteristic curves: Negative predictive value

Discussion

Video identification evidence is becoming increasingly commonplace in criminal investigations (Rumschik et al., 2020). In this thesis, I modeled face matching lineups after ones traditionally used for eyewitness identification. When fair filler-siphoning was assumed, face matching lineups were found to result in fewer estimated innocent suspect identifications and better PPV than 1-to-1 comparisons. Although this finding can only be generalized to circumstances where lineups are perfectly fair, this thesis represents a promising first step in the development of a reliable procedure for video evidence identification.

The current study revealed that face matching lineups can result in better PPV than 1-to-1 comparisons. However, this was only true under certain assumptions. When I classified all false alarms in lineups as innocent suspect identifications (M = 59%), 1-to-1 comparisons resulted in significantly fewer innocent suspect identifications (M = 20%). Similarly, when I estimated innocent suspect identifications using the effective size correction, which involves dividing false alarms by the number of plausible lineup members, 1-to-1 comparisons still resulted in significantly fewer innocent suspect identifications. The average effective size in target-absent lineups was less than three, indicating that over half of the lineup members were classified as implausible options. As a result, the estimated rate of innocent suspect identifications for lineups reduced to 23% when the effective size correction was applied. However, when I applied the nominal size correction, which involves dividing false alarms by lineup size (6), then lineups were estimated to result in significantly fewer innocent suspect identifications (10%) than 1-to-1 comparisons. These findings indicate that for face matching lineups to have an advantage over 1-to-1 comparisons, they must be perfectly fair, which is not always feasible in the context of video evidence identification where suspects are often under investigation for appearance-based reasons. Therefore, in practice, face matching lineups may not always result in better suspect identification accuracy than 1-to-1 comparisons.

The advantage for fair lineups over 1-to-1 comparisons did also result in a trade-off. Specifically, the reduction in innocent suspect identifications found with fair lineups came at a cost for correct identifications. While there is evidence of lineups and showups resulting in similar rates of correct identifications (Dysart & Lindsay, 2007; Gronlund et al., 2012; Steblay et al., 2003; Valentine et al., 2012), research has shown that lineups can result in fewer correct identifications than showups (Beal et al., 1995; Gonzales et al., 1993; Smith et al., 2023). Consistent with this finding, lineups in the current study resulted in significantly fewer correct identifications of the perpetrator (65%) compared to 1-to-1 comparisons (83%). As a result, it was only when innocent suspect identifications were sufficiently reduced through the nominal size correction that lineups resulted in better PPV than 1-to-1 comparisons. Furthermore, lineups also resulted in worse NPV than 1-to-1 comparisons. Therefore, although fair lineups have the potential to reduce innocent suspect identifications, 1-to-1 comparisons are better at apprehending guilty suspects and providing exculpatory information that can be used to rule out suspects.

Given that target-absent 1-to-1 comparisons were more often rejected than target-absent lineups, our findings are more consistent with differential filler-siphoning theory than diagnostic-feature detection theory. According to diagnostic-feature detection theory fair lineups are better than showups because they improve discrimination between guilty and innocent suspects (Wixted & Mickes, 2014). Contrarily, differential filler-siphoning theory argues that fair lineups are better than showups because fillers siphon more choices away from innocent suspects than guilty ones (Smith et al., 2017). In line with differential filler-siphoning theory, I found that it was only when I assumed perfect filler-siphoning (i.e., fair lineups) that lineups were superior to 1-to-1 comparisons in terms of PPV. However, the reduction in correct identifications was actually greater than the reduction in innocent suspect identifications estimated with the nominal size correction. Therefore, counter to what is proposed by differential filler-siphoning theory, this finding suggests that in the face matching context, fillers may actually siphon more choices away from guilty suspects than innocent ones. Note, however, that my lineups were not in fact perfectly fair, and the nominal size correction is only a method of estimating performance if the lineups were perfectly fair. This creates more uncertainty in whether fillers would siphon more choices from guilty suspects than innocent ones.

One explanation for this finding could be that face matching tasks elicit higher correct identification and lower innocent suspect identification rates than eyewitness tasks. Face matching can be easier than eyewitness identification because the latter relies on memory. For example, in the current study the correct identification rate for the

1-to-1 procedure was 83% and the innocent suspect identification rate was 20%. The addition of fillers led to a correct identification rate of 65% and an innocent suspect identification rate of 10%, when perfect fairness was assumed. By comparison, on an eyewitness identification task, Smith and colleagues (2017) found a 62% correct identification rate and 42% innocent suspect identification rate for showups versus a 68% correct identification rate and a 10% innocent suspect identification rate for lineups. Therefore, across these two studies, fillers had the potential to siphon more correct identifications on face matching tasks (83%) than eyewitness tasks (62%). Contrarily, fillers had the potential to siphon more innocent suspect identifications on eyewitness tasks (42%) than face matching tasks (20%). However, in a metanalysis on eyewitness identification, Steblay and colleagues (2003) found similar rates of innocent suspect identifications for showups (23%) and lineups (10%) as the current study. Therefore, avoiding the identification of an innocent suspect may not always be easier with face matching than with memory.

In practice, fillers in face matching lineups may have greater potential to siphon innocent suspect identifications. This is because contextual information present during real investigations can bias the identification task by creating expectations about whether the captured face matches the suspect's face (Saks et al., 2003). Research on forensic fingerprint analysis has shown that presenting prints alongside incriminating contextual information can increase false alarms (Quigley-McBride & Wells, 2018). The biasing effects of contextual information extend to face matching as well (Bruce et al., 2001). Given that video evidence identifications are typically made by police or other examiners within the police department (Facial Identification Scientific Working Group, 2022), it is likely that identifications are made in the presence of irrelevant contextual information. Furthermore, facial recognition technology often provides contextual information about match candidates, including similarity scores and, in some cases, arrest history (*United States of America v. Michael Joseph Peterson, Jr*., 2021). In a 1-to-1 comparison, the identity of the suspect is known; therefore, any incriminating information about the suspect will put them at a greater risk of being identified. This suggests that in real investigations, false alarms from 1-to-1 comparisons may be higher than those found in the current study. Consequently, in practice, face matching lineups could have greater potential to reduce innocent suspect identifications because, in addition to siphoning false alarms, they also conceal the identity of the suspect.

However, for face matching lineups to effectively protect against contextual bias they must be fair and there are currently no guidelines on how to construct fair face matching lineups.

The ROC analysis showed that the reduction in correct identifications and increase in misidentifications for lineups resulted in poorer investigator discriminability than 1-to-1 comparisons. This finding is consistent with previous research on lineups (Smith et al., 2017). Although the face matching lineups had worse investigator discriminability than the 1-to-1 comparisons, the rule-procedure boosted investigator discriminability in lineups to be more comparable to 1-to-1 comparisons. In line with previous research, this finding indicates that the rule-out is superior to standard lineups because of the additional information it provides, which can be useful for ruling out innocent suspects (Ayala et al., 2022). Therefore, if face matching lineups were to be adopted in practice, the rule-out procedure would likely be beneficial.

Consistent with previous face matching research (Bruce et al., 1999; Hahn et al., 2022; Stephens et al., 2017; White et al., 2014), I also found that confidence was predictive of accuracy for face matching. In the 1-to-1 condition, 10.3% of high confidence suspect identifications and 7.6% of high confidence rejections were errors. In the lineup condition, the error rate for high confidence suspect identifications again depended on how I estimated innocent suspect identifications. When all false alarms were assumed to be innocent suspect identifications, 29.4% of high confidence suspect identifications were errors. Additionally, when innocent suspect identifications were estimated using the effective size correction, 14.9% of high confidence suspect identifications were errors. However, when innocent suspect identifications were estimated using the nominal size correction, only 8.4% of high confidence suspect identifications were errors. Therefore, unless lineups are perfectly fair, confidence would not be a good indication of suspect guilt. Additionally, confidence was not a good indicator of accuracy for lineup rejections. In the lineup condition, 22.3% of high confidence rejections were errors. However, rule-out ratings were a better indication of accuracy, with 13.9% of high confidence rule-out ratings being errors. Together these results suggest that confidence can indicate accuracy for choices made from fair lineups that include the rule-out procedure or 1-to-1 comparisons.

While it is promising that face matching lineups can result in better PPV than 1-to-1 comparisons, this finding may not be applicable to real investigations. The nominal size correction assumes that innocent suspects are no more likely than lineup fillers to be mistakenly identified (Fitzgerald et al., 2023). Consequently, estimating innocent suspect identifications using the nominal size correction is most applicable to criminal cases where the identification conditions are pristine (Fitzgerald et al., 2023) and the innocent suspect is under investigation due to non-appearance-based reasons (Lee & Penrod, 2019). However, in the context of video evidence identification, it is likely that suspects are under investigation due to some level of resemblance to the perpetrator in the CCTV footage. This is especially true when facial recognition technology is used as an investigative tool to generate a list of candidates who are similar to the perpetrator in the capture. Therefore, unless fillers are also selected based on their resemblance to the person in the CCTV footage, innocent suspects will be at greater risk of being misidentified from video evidence. In support of this argument, when I considered lineup fairness and estimated innocent suspect identifications using the effective size correction, innocent suspects were actually at a greater risk of being misidentified from lineups than they were from 1-to-1 comparisons.

Furthermore, due to the trade-off between correct and innocent suspect identifications, lineups may not always produce better investigative outcomes than 1-to-1 comparisons. The cost of an identification outcome is the discrepancy between that outcome and the goal of the police investigation (Yang et al., 2019). In eyewitness identification research, it has been argued that lineups always produce a lower expected cost than showups. This is because the use of lineups in place of showups has not been found to result in a trade-off (Clark, 2012; Yang et el. 2019). However, I found that, in addition to reducing innocent suspect identifications (when I applied the nominal size correction), lineups also reduced correct identifications relative to 1-to-1 comparisons. Therefore, if a criminal justice system is more concerned with apprehending guilty suspects than protecting innocent ones, then the 1-to-1 comparison for face matching would be preferred. On the other hand, face matching lineups would be preferred in a justice system based on values that reflect Blackstone’s (1769, p.353) view that incriminating innocent suspects is disproportionately more costly than failing to incriminate guilty ones.

The impact of this trade-off also depends on the prior probability of guilt. If the prior probability of guilt is low, there will be a greater risk of misidentifying an innocent suspect, making fair face matching lineups the better option. However, if the prior probability of guilt is high, then the 1-to-1 method would likely be less costly. In the real world, prior probability of guilt is challenging to assess, and it varies between jurisdictions depending on the policies in place (Juncu & Fitzgerald, 2021; Wells & Olson, 2003).

One thing that can affect the prior probability of guilt is whether there is evidence-based suspicion before conducting an identification procedure. Evidence-based suspicion refers to evidence that leads to a reasonable conclusion that a certain individual, to the exclusion of all others, likely committed the crime (Wells et al., 2020). In the context of eyewitness identification, fitting a witness's general description of the perpetrator is not evidence-based suspicion because this description can be applied to many people (Wells et al., 2020). In a similar vein, having similarities to the perpetrator in the CCTV footage does not imply evidence-based suspicion. However, given that police typically start with CCTV footage, suspects are often only under investigation because someone or something (i.e., facial recognition technology) has decided that they look similar to the perpetrator in the capture. Because police often conduct 1-to-1 comparisons before having evidence-based grounds to investigate, prior probability of guilt is likely low for most video evidence identifications. Consequently, innocent people are at risk of being misidentified from video evidence. Therefore, if face matching lineups can be constructed fairly, then they should be used.

Implementing face matching lineups in practice could be challenging. Suspects are often under investigation because police officers recognize the individual in the CCTV footage (Keefe, 2016) or facial recognition technology has flagged them as a potential match candidate. In such cases, it is not possible to implement face matching lineups for the initial video evidence identification where an individual first becomes a suspect. However, even in such cases, face matching lineups could be used in the verification stage of the ACE-V method that FISWG recommend for facial examinations (Facial Identification Scientific Working Group, 2023). The verification stage involves a second independent examiner completing the analysis, comparison, and evaluation steps after a first examiner has already completed these steps and reached their own conclusion. This means that even if a suspect was initially identified through recognition, whereby some type of stored information from the suspect's face is retrieved from memory (Bruce & Young, 1986), or 1-to-1 comparison, a face matching lineup can be used with a second examiner to verify the initial conclusion. In doing so, face matching lineups may help to mitigate the biasing effects of the procedure used during the first identification and any contextual information that could have been present (Quigley-McBride & Wells, 2018). Furthermore, the use of face matching lineups during verification can also provide important exculpatory evidence. Specifically, if a second examiner does not pick a suspect that was initially identified through recognition or 1-to-1 comparison, then this may show evidence of that person's innocence.

The use of face matching lineups in practice may also help to identify unreliable police and examiners. The inclusion of fillers in face matching lineups means that this identification procedure is a test that can be failed (Wells et al., 2013). Unlike proficiency testing, where police and examiners are asked to identify individuals of known identity, face matching lineups can provide a measure of accuracy in actual cases. This could help to identify police or examiners who consistently choose fillers in actual cases.

Limitations

Although these findings represent a promising first step in the development of a procedure that improves video evidence identification, my study has limitations to consider. First, the experiments were completed online in a single session lasting 20-30 minutes. Although video evidence identifications in criminal investigations could also be made from computer screens, the context in which participants completed the study likely differed, nonetheless. Furthermore, administering the study online limited my ability to ensure participants understood task instructions. Second, the sample consisted of undergraduate students rather than professionals, such as police, forensic face examiners, or super-recognizers. Although, police have been found to perform no better than untrained participants on face matching tasks (Wirth & Carbon, 2017), there is evidence suggesting that forensic face examiners (Towler et al., 2017; Phillips, et al., 2018) and super-recognizers (Davis et al., 2016; Robertson et al., 2016) are better than average at face matching. Therefore, accuracy in the current study may be worse than it is in practice. Third, participants only saw a capture from a crime video footage. In real criminal investigations, police and examiners can watch and analyze the CCTV footage. Watching CCTV footage can provide additional information, such as gait, which may be useful for identification (Birch et al., 2015). Fourth, the capture showed the perpetrator head-on. To cover large areas, CCTV cameras are often placed high up (Rogers, 2019). As a result, CCTV footage can show perpetrators from odd angles, which can impair face matching accuracy (Bruce et al., 1999). Lastly, participants knew no crime took place, which can affect identification decisions (Eisen et al., 2022).

Future Directions

My findings highlight the need for more research on face matching lineups. Although I demonstrated that face matching lineups can be better than 1-to-1 comparisons, the generalizability of this finding is limited to pristine identification conditions. Therefore, future research is needed to better understand how face matching lineups should be constructed in practice.

One avenue for future research would be to explore how fillers should be selected. For eyewitness identification, it is recommended that lineups should contain appropriate fillers who match the description of the perpetrator and do not make the suspect stand out (Wells, 2020). For face matching lineups, it is unclear how fillers should be selected, especially when someone becomes a suspect via facial recognition technology. If a suspect is under investigation because facial recognition technology selected them as a match candidate, then they will likely have high similarity to the perpetrator in the captured image. In such circumstances, it could be challenging to find plausible fillers who are as similar to the perpetrator as the suspect. One possible solution for generating plausible fillers is to submit the captured image to facial recognition technology that is being used in a different location (Wells, 2024). Doing so would produce a list of match candidates who are known to be innocent because they were in a different location at the time of the crime. This could enable the construction of fair face matching lineups; however, further research is needed to examine the validity and feasibility of this method.

Additional future research should investigate the ideal size for face matching lineups. Following Wells and colleagues' (2020) recommendation for eyewitness identification lineups, I used six-person lineups in the current study. However, eyewitness identification lineup size varies across jurisdictions, ranging from lineups as small as three to lineups as large as 10 (Fitzgerald et al., 2021). Research has shown that while larger lineups provide better protection to innocent suspects than smaller lineups, they do make it more challenging for witnesses to identify guilty suspects (Juncu & Fitzgerald, 2021). The optimal number of fillers for face matching lineups is still unknown and should therefore be investigated in future research.

Lastly, future research should use more realistic conditions to address the limitations of this study. First, a replication of this study should be attempted with professional samples, including police, forensic face examiners, and super-recognizers. Second, future research should examine the effects of contextual information on video evidence identification and determine whether face matching lineups can help mitigate biases that arise. Third, to improve the external validity of our findings, face matching lineups should be tested with CCTV footage that participants can watch and pause. Similarly, face matching lineups should also be tested with CCTV images that show the perpetrator's face from odd angles and in poor resolution.

Conclusion

In this thesis, I tested face matching lineups in a video evidence identification task. With the filler-control method, I found that under pristine conditions face matching lineups can result in better identification outcomes than 1-to-1 comparisons. In line with differential filler-siphoning theory, my results demonstrate that face matching lineups are only superior to 1-to-1 comparisons because of filler-siphoning. Although face matching lineups are yet to be perfected, the findings from this thesis present a promising solution to the risk of misidentification from video evidence.

References

Armanto, T. (1997). Snake [Computer software]. Findland: Rumilus Design

Attorney General's Reference. [No.2 of 2002], (2003). 1 Cr. App. R. 321, England.

Bate, S., Frowd, C., Bennetts, R., Hasshim, N., Murray, E., Bobak, A. K., Wills, H., & Richards, S. (2018). Applied screening tests for the detection of superior face recognition. *Cognitive Research: Principles and Implications*, *3*(1). https://doi.org/10.1186/s41235-018-0116-5

Blackstone, W. (1769). *Commentaries on the laws of England: Vol. II, Book IV*. Duyckinck, Long, Collins & Hannay, and Collins & Co.

Bindemann, M., Attard, J., & Johnston, R. A. (2014). Perceived ability and actual recognition accuracy for unfamiliar and famous faces. *Cogent Psychology, 1*(1). https://doi.org/10.1080/23311908.2014.986903

Birch, I., Birch, M., & Asgeirsdottir, N. (2020). The identification of individuals by observational gait analysis using closed circuit television footage: Comparing the ability and confidence of experienced and non-experienced analysts. *Science & Justice, 60*(1), 79-85.

Bobak, A. K., Hancock, P. J., & Bate, S. (2016). Super-recognisers in action: Evidence from face-matching and face memory tasks. *Applied Cognitive Psychology, 30*(1), 81–91. https://doi.org/10.1002/acp.3170

Brigham, J. C., & Malpass, R. S. (1985). The role of experience and contact in the recognition of faces of own- and other-race persons. *Journal of Social Issues, 41*(3), 139–155. https://doi.org/10.1111/j.1540-4560.1985.tb01133.x

Bruce, V., & Young, A. (1986). Understanding face recognition. *British Journal of Psychology, 77*, 305–327. https://doi.org/10.1111/j.2044-8295.1986.tb02199

Bruce, V., Henderson, Z., Greenwood, K., Hancock, P., Burton, A., & Miller, P. (1999). Verification of face identities from images captured on video. *Journal of Experimental Psychology: Applied*, 5(4), 339-360. https://doi.org/10.1037/1076-898x.5.4.339

Bruce, V., Henderson, Z., Newman, C., & Burton, A. M. (2001). Matching identities of familiar and unfamiliar faces caught on CCTV images. *Journal of Experimental Psychology: Applied, 7*(3), 207–218. https://doi.org/10.1037/1076-898X.7.3.207

Bhuiyan, J. (2023). *"Are you kidding, carjacking?": The problem with facial recognition in policing*. The Guardian. https://www.theguardian.com/newsletters/2023/aug/15/techscape-facial-recognition-software-detroit-porcha-woodruff-black-people-ai

Burton, A. M., White, D., & McNeill, A. (2010). The Glasgow Face Matching Test. *Behavior Research Methods, 42*(1), 286–291. https://doi.org/10.3758/brm.42.1.286

Burton, A. M., Wilson, S., Cowan, M., & Bruce, V. (1999). Face recognition in poor-quality video: Evidence from security surveillance. *Psychological Science, 10*(3), 243–248. https://doi.org/10.1111/1467-9280.00144

Carragher, D.J., Hancock, P.J.B. (2020). Surgical face masks impair human face matching performance for familiar and unfamiliar faces. *Cognitive Research: Principles and Implications 5*(59), 1-15. https://doi.org/10.1186/s41235-020-00258-x

Carragher, D. J., & Hancock, P. J. B. (2023). Simulated automated facial recognition systems as decision-aids in forensic face matching tasks. *Journal of Experimental Psychology: General, 152*(5), 1286–1304. https://doi.org/10.1037/xge0001310

Clark, S. E., & Godfrey, R. D. (2009). Eyewitness identification evidence and innocence risk. *Psychonomic Bulletin & Review, 16*(1), 22–42. https://doi.org/10.3758/pbr.16.1.22

Clark, S. E. (2012). Costs and benefits of eyewitness identification reform: Psychological science and public policy. *Perspectives on Psychological Science, 7(*3), 238-259.

Clutterbuck, R., & Johnston, R. A. (2002). Exploring levels of face familiarity by using an indirect face-matching measure. *Perception*, *31*(8), 985-994. https://10.1068/p3335

Colloff, M. F., Wade, K. A., & Strange, D. (2016). Unfair lineups make witnesses more likely to confuse innocent and guilty suspects. *Psychological Science*, *27*(9), 1227-1239.

Cross, J. F., Cross, J., & Daly, J. (1971). Sex, race, age, and beauty as factors in recognition of faces. *Perception & Psychophysics*, *10*(6), 393-396. https://10.3758/bf03210319

Davis, J. P., & Valentine, T. (2009). CCTV on trial: Matching video images with the defendant in the dock. *Applied Cognitive Psychology, 23*(4), 482-505. https://10.1002/acp.1490

Davis, J. P., Lander, K., Evans, R., & Jansari, A. (2016). Investigating predictors of superior face recognition ability in police super-recognisers. *Applied Cognitive Psychology, 30*, 827–840. doi:10.1002/acp.3260

Davidovic, I. (2019, November 18). *Should we be worried by ever more CCTV cameras?* BBC News. Retrieved December 7, 2022, from https://www.bbc.com/news/business-50348861

Dhamecha, T. I., Singh, R., Vatsa, M., & Kumar, A. (2014). Recognizing disguised faces: Human and machine evaluation. *PLoS ONE, 9*(7), e99212. https://doi.org/10.1371/journal.pone.0099212

Dystart, J. E., & Lindsay, R. C. L. (2007). Show-up identifications: Suggestive technique or reliable method? . In Lindsay, R.C.L., Ross, D.F., Read, J.D., & M. P. Toglia, (Eds.), *Handbook of Eyewitness Psychology: Volume II* (pp. 151–168). PSYCHOLOGY Press.

Eisen, M. L., Ying, R. C., Chui, C., & Swaby, M. A. (2022). Comparing witness performance in the field versus the lab: How real-world conditions affect eyewitness decision-making. *Law and Human Behavior, 46*(3), 175–188. https://doi.org/10.1037/lhb0000485

Ellis, H. D., Shepherd, J. W., & Davies, G. M. (1979). Identification of familiar and unfamiliar faces from internal and external features: Some implications for theories of face recognition. *Perception, 8*(4), 431–439. https://doi.org/10.1068/p080431

Estudillo, A. J., Hills, P., & Wong, H. K. (2021). The effect of face masks on forensic face matching: An individual differences study. *Journal of Applied Research in Memory and Cognition, 10*(4), 554–563.

Facial Identification Scientific Working Group. (2022). Minimum Guidelines for Facial Image Comparison Documentation. https://fiswg.org/fiswg_min_guidelines_for_facial_image_comp_docum_v2.0_2022.11.04.pdf

Facial Identification Scientific Working Group. (2023). ACE-V Methodology for the Use in One-to-One Examinations Document. https://fiswg.org/FISWG_ACE-V%20methodology%20for%20the%20Use%20in%20One-to-One%20Examinations_V1.0_20231117.pdf

Facial Identification Scientific Working Group. (n.d). *About Scientific Working Groups*. https://www.fiswg.org/about.html

Finklea, K., Harris, L.A., Kolker, A.F., & Sargent, J.F. (2023). Federal law enforcement use of facial recognition technology. https://crsreports.congress.gov/product/pdf/R/R46586

Fitzgerald, R. J., Rubínová, E., & Juncu, S. (2021). Eyewitness identification around the world. In A. M. Smith, M. P. Toglia, & J. M. Lampinen (Eds.), Methods, measures, and theories in eyewitness identification tasks (pp. 294–322). Routledge, Taylor & Francis Group.

Fitzgerald, R. J., Tredoux, C.G., & Juncu, S. (2023). Estimation of eyewitness error rates in fair and biased lineups. *Law and Human Behavior, 47*(4), 463-483.

Flowe, H. D., Klatt, T., & Colloff, M. F. (2014). Selecting fillers on emotional appearance improves lineup identification accuracy. Law and Human Behavior, 38(6), 509-519.

Forensic Science Regulator (2016). Codes of Practice and Conduct for forensic science providers and practitioners, 3. Available at https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/499850/2016_2_11_-_The_Codes_of_Practice_and_Conduct_-_Issue_3.pdf

Garrett, B. L., Liu, A., Kafadar, K., Yaffe, J., & Dodson, C. S. (2020). Factoring the role of eyewitness evidence in the courtroom. *Journal of Empirical Legal Studies*, *17*(3), 556-579. https://10.1111/jels.12259

Gill, M., & Loveday, K. (2003). What do offenders think about CCTV?. *Crime prevention and community safety*, *5*(3), 17-25. https://10.1057/palgrave.cpcs.8140152

Goodsell, C. A., Wetmore, S. A., Neuschatz, J. S., & Gronlund, S. D. (2013). Showups. In B. L. Cutler (Ed.), *Reform of eyewitness identification procedures* (pp. 45–63). American Psychological Association. https://doi.org/10.1037/14094-003

Gonzalez, R., Ellsworth, P. C., & Pembroke, M. (1993). Response biases in lineups and showups. *Journal of personality and Social Psychology, 64*(4), 525–537.

Graham, D. L., & Ritchie, K. L. (2019). Making a spectacle of yourself: The effect of glasses and sunglasses on face perception. *Perception, 48*(6), 461–470.

Gronlund, S. D., Carlson, C. A., Neuschatz, J. S., Goodsell, C. A., Wetmore, S. A., Wooten, A., & Graham, M. (2013). Showups versus lineups: An evaluation using ROC analysis. *Journal of Applied Research in Memory and Cognition*, *1*(4), 221-228. . https://10.1037/e571212013-331

Hahn, C. A., Tang, L. L., Yates, A. N., & Phillips, P. J. (2022). Forensic facial examiners versus super-recognizers: Evaluating behavior beyond accuracy. *Applied Cognitive Psychology*, 36(6), 1209–1218. https://doi.org/10.1002/acp.4003

Henderson, Z., Bruce, V., & Burton, A. M. (2001). Matching the faces of robbers captured on video. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, *15*(4), 445-464. https://10.1002/acp.718

Hockley, W. E., Hemsworth, D. H., & Consoli, A. (1999). Shades of the mirror effect: Recognition of faces with and without sunglasses. *Memory & Cognition*, *27*(1), 128-138. https:// 10.3758/bf03201219

Innocence Project. (2010). *Race and wrongful convictions*. www.innocenceproject.org/race-and-wrongful-convictions/

Juncu, S. & Fitzgerald, R.J. (2021). A meta-analysis of lineup size effects on eyewitness identification. Psychology, Public Policy and Law, 27(3), 295-315. https://doi.org/10.1037/law0000311

Kassin, S. M., Tubb, V. A., Hosch, H. M., & Memon, A. (2001). On the “general acceptance” of eyewitness testimony research. *American Psychologist, 56*, 405–416.

Kassin, S.M., Dror, I.D., Kukucka, J. (2013). The forensic confirmation bias: Problems, perspectives, and proposed solutions. *Journal of Applied Research in Memory and Cognition, 2*(1), 42-52. https://doi.org/10.1016/j.jarmac.2013.01.001

Keefe, P. R. (2016). *The detectives who never forget a face*. The New Yorker. https://www.newyorker.com/magazine/2016/08/22/londons-super-recognizer-police-force

Kemp, R., Towell, N., & Pike, G. (1997). When seeing should not be believing: Photographs, credit cards and fraud. *Applied Cognitive Psychology*, *11*(3), 211-222. https:// 10.1002/(sici)1099-0720(199706)11:3<211::aid-acp430>3.0.co;2-o

Kramer, R. S., & Ritchie, K. L. (2016). Disguising superman: How glasses affect unfamiliar face matching. *Applied Cognitive Psychology*, *30*(6), 841-845. https:// 10.1002/acp.3261

Lee, J., & Penrod, S. D. (2019). New signal detection theory-based framework for eyewitness performance in lineups. *Law and Human Behavior, 43(*5), 436–454. https://doi.org/10.1037/lhb0000343

Lee, W. J., Wilkinson, C., Memon, A., Houston, K., & Res, M. (2009). Matching Unfamiliar Faces from Poor Quality Closed-Circuit Television (CCTV) Footage: An evaluation of the effect of training on facial identification ability. *AXIS, 1*(1), 19-28.

Lindsay, R. C., & Wells, G. L. (1985). Improving eyewitness identifications from lineups: Simultaneous versus sequential lineup presentation. *Journal of Applied Psychology, 70*(3), 556-564. https:// 10.1037/0021-9010.70.3.556

Lynch, J. (2020). Face off: Law enforcement use of face recognition technology. https://ssrn.com/abstract=3909038 or http://dx.doi.org/10.2139/ssrn.

Malpass, R. S. (1981). Effective size and defendant bias in eyewitness identification lineups. *Law and Human Behavior, 5*(4), 299–309. https://doi.org/10.1007/BF01044945Megreya,

Megreya, A. M., & Burton, A. M. (2006). Unfamiliar faces are not faces: Evidence from a matching task. *Memory & Cognition*, *34*(4), 865-876. https:// 10.3758/bf03193433

Megreya, A. M., & Burton, A. M. (2007). Hits and false positives in face matching: A familiarity-based dissociation. *Perception & Psychophysics*, *69*(7), 1175-1184. https:// 10.3758/bf03193954

Megreya, A. M., & Burton, A. M. (2008). Matching faces to photographs: poor performance in eyewitness memory (without the memory). *Journal of Experimental Psychology: Applied*, *14*(4), 364-372. https://doi.org/10.1037/a0013464

Megreya, A. M., White, D., & Burton, A. M. (2011). The other-race effect does not rely on memory: Evidence from a matching task. *Quarterly Journal of Experimental Psychology*, *64*(8), 1473-1483. https:// 10.1080/17470218.2011.575228

Megreya, A. M., Sandford, A., & Burton, A. M. (2013). Matching face images taken on the same day or months apart: The limitations of photo ID. *Applied Cognitive Psychology*, *27*(6), 700-706. https:// 10.1002/acp.2965

Meissner, C. A., & Brigham, J. C. (2001). Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychology, Public Policy, and Law*, *7*(1), 3-35. https:// 10.1037/1076-8971.7.1.3

Mickes, L. (2015). Receiver operating characteristic analysis and confidence-accuracy characteristic analysis in investigations of system variables and estimator variables that affect eyewitness memory. *Journal of Applied Research in Memory and Cognition, 4*(2), 93-102.

Moreton, R. (2021). Forensic Face Matching: Procedures and Application. In M. Bindemann

(Ed.), *Forensic face matching: Research and practice* (pp. 144–173). Oxford University Press. https://doi.org/10.1093/oso/9780198837749.003.0007

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

Neauschatz, J. N., Wetmore, S. A., Key, K. N., Cash, D. K., Gronlund, S. D., & Goodsell, C. A. (2016). A comprehensive evaluation of showups. In M. K. Miller, B. H. Bornstein, & D. DeMatteo (Eds.), *Advances in psychology and law* (Vol. 1, pp. 43–69). Springer.

Nosworthy, G. J., & Lindsay, R. C. (1990). Does nominal lineup size matter?. *Journal of Applied Psychology*, *75*(3), 358-361. https:// 10.1037/0021-9010.75.3.358

Office of the Privacy Commissioner of Canada (2022). *Privacy guidance on facial recognition for police agencies*. https://www.priv.gc.ca/en/privacy-topics/surveillance/police-and-public-safety/gd_fr_202205/

O'Toole, A. J., Deffenbacher, K. A., Valentin, D., & Abdi, H. (1994). Structural aspects of face recognition and the other-race effect. *Memory & Cognition*, *22*(2), 208-224. https:// 10.3758/bf03208892

O'Toole, A. J., An, X., Dunlop, J., Natu, V., & Phillips, P. J. (2012). Comparing face recognition algorithms to humans on challenging tasks. *ACM Transactions on Applied Perception*, *9*(4), 1–13. https://doi.org/10.1145/2355598.2355599

Palmer, M.A., Brewer, N. (2012). Sequential lineup presentation promotes less-biased criterion setting but does not improve discriminability. *Law and Human Behavior, 36*, 247-255. http://dx.doi.org/10.1037/h0093923

Phillips, P. J., Yates, A. N., Hu, Y., Hahn, C. A., Noyes, E., Jackson, K., & O'Toole, A. J. (2018). Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. *Proceedings of the National Academy of Sciences*, *115*(24), 6171-6176. https:// 10.1073/pnas.1721355115

Quigley-McBride, A., & Wells, G. L. (2018). Fillers can help control for contextual bias in forensic comparison tasks. Law and Human Behavior, 42(4), 295–305. https://doi.org/10.1037/lhb0000295

Ritchie, K. L., Carragher, D. J., Davis, J. P., Read, K., Jenkins, R. E., Noyes, E., Gray, L.H., & Hancock, P. J. (2024). Face masks and fake masks: the effect of real and superimposed masks on face matching with super-recognisers, typical observers, and algorithms. *Cognitive Research: Principles and Implications, 9*(1), 1-13. https://doi.org/10.1186/s41235-024-00532-2

Robertson, D. J., Noyes, E., Dowsett, A. J., Jenkins, R., & Burton, A. M. (2016). Face recognition by Metropolitan Police super-recognisers. *PLoS ONE, 11*(2), e0150036. doi:10.1371/journal.pone.0150036

Rogers, S. (2019). Why are CCTV images still so rubbish?. BBC Science Focus Magazine - Science, Nature, technology, q&as. https://www.sciencefocus.com/science/why-are-cctv-images-still-so-rubbish

Rowley, S. (2009). *Wrongful conviction throws spotlight on unreliability of eyewitness evidence*. The Guardian. Retrieved from https://www.theguardian.com/uk/2009/aug/18/eyewitness-evidence-wrongful-conviction

Rumschik, D. M., Berman, G. L., & Cutler, B. L. (2020). Person-Matching: Real-Time

Identifications of Persons from Photos and Videos. In M. K. Miller & B. H. Bornstein (Eds.), *Advances in Psychology and Law: Volume 5* (pp. 1–22). Springer International Publishing. https://doi.org/10.1007/978-3-030-54678-6_1

Saks, M. J., Risinger, D. M., Rosenthal, R., & Thompson, W. C. (2003). Context effects in forensic science: A review and application of the science of science to crime laboratory practice in the United States. *Science & Justice, 43*, 77–90.

Sanford, A. (2024). *Artificial Intelligence is putting innocent people at risk of being incarcerated*. Innocence Project. https://innocenceproject.org/artificial-intelligence-is-putting-innocent-people-at-risk-of-being-incarcerated/

Sandford, A., & Ritchie, K. L. (2021). Unfamiliar face matching, within-person variability, and multiple-image arrays. *Visual Cognition*, *29*(3), 143-157. https:// 10.1080/13506285.2021.1883170

Shapiro, P. N., & Penrod, S. (1986). Meta-analysis of Facial Identification Studies. *Psychological Bulletin, 100*(2), 139–156. https://doi.org/10.1037/0033-2909.100.2.139

Smith, A. M., Smalarz, L., Ditchfield, R., & Ayala, N. T. (2021). Evaluating the claim that high confidence implies high accuracy in eyewitness identification. *Psychology, Public Policy, and Law, 27*(4), 479–491. https://doi.org/10.1037/law0000324

Smith, A. M., Wells, G. L., Lindsay, R. C. L., & Penrod, S. D. (2017). Fair lineups are better than biased lineups and showups, but not because they increase underlying discriminability. *Law and Human Behavior, 41*, 127-145. http://dx.doi.org/10.1037/lhb0000219

Smith, A. M., Yang, Y., & Wells, G. L. (2020). Distinguishing between investigator discriminability and eyewitness discriminability: A method for creating full receiver operating characteristic curves of lineup identification performance. *Perspectives on Psychological Science, 15*(3), 589–607. https://doi.org/10.1177/1745691620902426

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

Smith, A. M., Ying, R. C., Goldstein, A. R., & Fitzgerald, R. J. (under review). Absolute-judgment models better predict eyewitness decision-making than do relative-judgment models.

Steblay, N., Dysart, J., Fulero, S., & Lindsay, R.C.L. (2003). Eyewitness accuracy rates in police showup and lineup presentations: A meta-analytic comparison. *Law and Human Behavior, 27*, 523-540. https://doi.org/10.1023/A:1025438223608

Steblay, N. K. (2019). Translating psychological science into policy and practice. In N. Brewer & A. B. Douglass (Eds.), Psychological science and the law (pp. 417–443). The Guilford Press.

Stephens, R. G., Semmler, C., & Sauer, J. D. (2017). The effect of the proportion of mismatching trials and task orientation on the confidence–accuracy relationship in unfamiliar face matching. *Journal of Experimental Psychology: Applied, 23*(3), 336.

Susa, K. J., Michael, S. W., Dessenberger, S. J., & Meissner, C. A. (2019). Imposter identification in low prevalence environments. *Legal and Criminological Psychology*, *24*(1), 179-193. https://doi.org/10.1111/lcrp.12138

Terry, R. L. (1993). How wearing eyeglasses affects facial recognition. *Current Psychology*, *12*(2), 151-162. https:// 10.1007/bf02686820

Towler, A., White, D., & Kemp, R. I. (2017). Evaluating the feature comparison strategy for forensic face identification. *Journal of Experimental Psychology: Applied, 23*, 47–58. doi:10.1037/xap0000108

Towler, A., Kemp, R. I., Burton, A. M., Dunn, J. D., Wayne, T., Moreton, R., et al. (2019). Do professional facial image comparison training courses work? *PLoS ONE, 14*(2), e0211037. doi:10.1371/journal.pone.0211037

Tredoux, C. G. (1998). Statistical inference on measures of lineup fairness. *Law and Human Behavior, 22*(2), 217–237. https://doi.org/10.1023/A1025746220886

Valentine, T., & Bruce, V. (1986). The effect of race, inversion and encoding activity upon face recognition. *Acta psychologica*, *61*(3), 259-273. https:// 10.1016/0001-6918(86)90085-5

Walker, P. M., & Hewstone, M. (2006). A developmental investigation of other-race contact and the own-race face effect. *British Journal of Developmental Psychology*, *24*(3), 451-463. https:// 10.1348/026151005x51239

Wells, G. L., Lindsay, R. C. (1980). On estimating the diagnosticity of eyewitness nonidentifications. *Psychological Bulletin, 88*(3), 776–784. https://doi.org/10.1037/0033-2909.88.3.776

Wells, G. L., & Turtle, J. W. (1986). Eyewitness identification: The importance of lineup models. *Psychological bulletin*, *99*(3), 320-329. https:// 10.1037/0033-2909.99.3.320

Wells, G. L. (2001). Police lineups: Data, theory, and policy. *Psychology, Public Policy, and Law, 7*(4), 791–801. https:// 10.1037/1076-8971.7.4.791

Wells, G.L., & Olson, E.A. (2003). Eyewitness testimony. *Annual Review of Psychology, 54*, 277-295. https://doi.org/10.1146/annurev.psych.54.101601.145028

Wells, G. L., & Penrod, S. D. (2011). Eyewitness identification research: Strengths and weaknesses of alternative methods. In B. Rosenfeld & S. D. Penrod (Eds.), *Research methods in forensic psychology* (pp. 237–256). Wiley.

Wells, G. L., Steblay, N. K., & Dysart, J. E. (2012). Eyewitness identification reforms: Are suggestiveness-induced hits and guesses true hits? *Perspectives on Psychological Science, 7*, 264 –271. http://dx.doi.org/10.1177/1745691612443368

Wells, G. L., Wilford, M.M., & Smalarz, L. (2013). Forensic science testing: The forensic filler-control method for controlling contextual bias, estimating error rates, and calibrating analysts' reports. *Journal of Applied Research in Memory and Cognition, 2*, 53-55. http://dx.doi.org/10.1016/j.jarmac.2013.01.004

Wells, G. L., Smalarz, L., & Smith, A. M. (2015). ROC analysis of lineups does not measure underlying discriminability and has limited value. *Journal of Applied Research in Memory and Cognition*, *4*(4), 313-317. https:// 10.1016/j.jarmac.2015.08.008

Wells, G. L., Steblay, N. K., & Dysart, J. E. (2015). Double-blind photo lineups using actual eyewitnesses: an experimental test of a sequential versus simultaneous lineup procedure. *Law and Human Behavior, 39*(1), 1-14. https:// 10.1037/lhb0000096

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

Wells, G. L. (2024, March 23). A nascent concern for eyewitness experts: Face-recognition technology contributes to high-confidence mistaken identifications. [Individual paper presentation]. American Psychology-Law Society Annual Conference, Los Angeles, C.A.

Wetmore, S. A., McAdoo, R. M., Gronlund, S. D., & Neuschatz, J. S. (2017). The impact of fillers on lineup performance. *Cognitive Research: Principles and Implications, 2*(1), 48. https://doi.org/10.1186/s41235-017-0084-1

Wirth, B. E., & Carbon, C. C. (2017). An easy game for frauds? Effects of professional experience and time pressure on passport-matching performance. *Journal of Experimental Psychology: Applied, 23*(2), 138–157. https://doi.org/10.1037/xap0000114

Wixted, J. T., & Mickes, L. (2012). The field of eyewitness memory should abandon probative value and embrace receiver operating characteristic analysis. *Perspectives on Psychological Science, 7*(3), 275-278

Wixted, J. T., & Mickes, L. (2014). A signal-detection-based diagnostic-feature-detection model of eyewitness identification. *Psychological Review, 121*(2), 262–276. https://doi.org/10.1037/a0035940

Wixted, J. T., & Wells, G. L. (2017). The relationship between eyewitness confidence and identification accuracy: A new synthesis. *Psychological Science in the Public Interest, 18*(1), 10-65.

White, D., Kemp R.I., Jenkins, R., Matheson, M., & Burton, A.M. (2014) Passport Officers' Errors in Face Matching. *PLoS ONE, 9*(8), 1-6. doi:10.1371/journal.pone.0103510

Wooten, A. R., Carlson, C. A., Lockamyeir, R. F., Carlson, M. A., Jones, A. R., Dias, J. L., & Hemby, J. A. (2020). The number of fillers may not matter as long as they all match the description: The effect of simultaneous lineup size on eyewitness identification. *Applied Cognitive Psychology, 34*(3), 590–604. https://doi.org/10.1002/acp.3644

Yang, Y., Smalarz, L., Moody, S. A., Cabell, J. J., & Copp, C. J. (2019). An expected cost model of eyewitness identification. *Law and Human Behavior, 43*(3), 205–219. https://doi.org/10.1037/lhb0000331

Yang, Y., & Smith, A. M. (2022). FullROC: An R package for generating and analyzing eyewitness-lineup ROC curves. *Behavior Research Methods, 55*(3), 1259–1274. https://doi.org/10.3758/s13428-022-01807-6

Young, A. W., Hay, D. C., McWeeny, K. H., Flude, B. M., & Ellis, A. W. (1985). Matching familiar and unfamiliar faces on internal and external features. *Perception, 14*(6), 737-746. https:// 10.1068/p140737

Cases

United States of America v. Michael Joseph Peterson, Jr., 25 U.S. 9 (2021).