UNIVERSITY OF BIRMINGHAM

OPTIMIZING LINEUP CONSTRUCTION: THE IMPACT OF FILLER SIMILARITY, DISTINCTIVE FACIAL FEATURES, AND INDIVIDUAL DIFFERENCES IN FACIAL RECOGNITION ON LINEUP PERFORMANCE

By

Georgia Roughton

A thesis submitted to the University of Birmingham

for the degree of

Doctorate in Forensic Psychology Practice (ForenPsyD)

Centre of Applied Psychology

School of Psychology

College of Life and Environmental Sciences

University of Birmingham

August 2022

Abstract

The present thesis sought to investigate optimal lineup construction methods that enhance ability to discriminate between innocent and guilty suspects. Firstly, I highlight the need for lineup construction methods that enhance eyewitness discriminability. Next, in chapter 2, I conducted a systematic literature review of suspect-filler similarity and found that there were no standardised procedures for constructing lineups in experiments. I also highlight the impact of methodological factors on discriminability. Using the feature matching model and diagnostic feature detection theory (Colloff et al., 2021; Wixted & Mickes, 2014), I argue that low similarity lineups allow the witness to focus on the perpetrators' unique features that are diagnostic of guilt, to make an accurate identification decision. However, I note that the low similarity lineup advantage holds only when lineup construction methods are fair (i.e., result in two memory strength distributions in witness memory: one for the perpetrator and one for the fillers and innocent suspect). In chapter 3, I conducted an experiment investigating lineup construction methods for distinctive suspects (e.g., with a facial tattoo). I compared a high similarity replication lineup, in which the distinctive suspects' facial tattoo is exactly replicated across lineup members; a low similarity replication lineup, in which the lineup members have a similar but non-identical distinctive facial tattoo; and a do nothing lineup in which only the suspect has a distinctive feature. As predicted by the feature matching model, I found that low similarity replication lineups yield higher discriminability compared to high similarity replication and do nothing lineups. In chapter 4, I critically evaluate the Benton Facial Recognition Test and advise on the use of psychometric tools to allow for further exploration of lineup construction methods that enhance discriminability when individual differences are also considered. Finally in chapter 5, I disentangle mixed findings in the literature to date and recommend that future research thoroughly reports lineup construction methods.

Acknowledgements

I would like to thank Dr Melissa Colloff for her continued support and guidance throughout the completion of this thesis. I continue to be inspired by Melissa's work, and I am truly grateful for the opportunity to lead on this project in such an active field of research. And further thanks to Dr Heather Flowe for overseeing my academic supervision in Melissa's absence, it has been a pleasure to work together.

Thank you to collaborator Aleena Mahmood for her contributions to stimuli creation and research development and to Tia Bennett for dual coding reviewed experiments. It was a pleasure to work with you both and I wish you all the best in your careers.

I would also like to express thanks to all the participants who were involved in the data collection and to the professionals I liaised with during the course of the thesis.

To all of my family and friends, thank you for your words of support and encouragement. Thank you for understanding when I had to prioritise study over spending time together, and for being patient for the day I finally finish studying! I could not have achieved any of this without you all and I am truly grateful.

To my Mum, thank you for showing me that helping others is a truly rewarding and fulfilling career and for inspiring me to pursue this career path.

To my Nan, thank you for being an anchor in my life and for taking care of me when it got tough. Thank you for being my biggest cheerleader and always believing in me.

To Stuart, thank you for listening, and bringing me happiness in the midst of a pandemic, new job, finishing a doctorate and a relocation! But most of all, thank you for being you.

Publications / Contributions

Research from Chapter 3 on "constructing lineups for distinctive suspects" was accepted to be presented at three conferences:

1. American Psychology-Law Society Annual Conference. This was presented as a paper in person by the author in Denver, USA (March 2022).
2. British Association of Cognitive Neuroscience. This was presented as a poster in person by the author in Birmingham, UK (May 2022).
3. British Psychological Society, Division of Forensic Psychology Annual Conference. This was presented as a paper in person by the author in Solihull UK (June 2022).

Research from Chapter 3 "constructing lineups for distinctive suspects" has also been accepted to be presented as a poster at the Psychonomic Society Annual Meeting 2022 and a paper is currently in preparation for publication.

Conference applications and presentation preparations were supported by Dr Melissa Colloff and Dr Heather Flowe, who also supervised this thesis.

During the experiment design phase, materials were developed in collaboration with Aleena Mahmood, a University of Birmingham MSci student who was also researching the effect of distinctive facial features.

For the systematic literature review, data were dual coded by second a second reviewer Tia Bennett and any discrepancies in coding was discussed with project supervisor Dr Melissa Colloff.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1 : INTRODUCTION

A witness to a crime may be later asked by authorities to identify the perpetrator from an array of people including the suspect of the crime, and this is known as a lineup. The witness may be faced with a lineup in which the perpetrator is present (i.e., the police suspect is the guilty person; target present) or in which the perpetrator is absent because the suspect is an innocent person (i.e., target absent). Also within the lineup, there will be other people who are known to be innocent (i.e., fillers). In experiments, this factor can be manipulated so that the presence or absence of the perpetrator is known by the researcher. And in target absent lineups, a designated innocent suspect can be presented alongside fillers (see Appendix G for a glossary of terminology used in this thesis). However, in real life, the guilt or innocence of the suspect is not known. Therefore, there is the risk that the witness may incorrectly identify a suspect as the perpetrator when they are actually innocent. And this problem can lead to wrongful conviction of innocent individuals (Quigley-McBride & Wells, 2021). For example, eyewitness misidentification was a factor involved in 69% of the 375 overturned convictions of innocent individuals by DNA evidence in the US to date (Innocence Project, 2022). Moreover, witnesses can also fail to identify the perpetrator of a crime when they are present within the lineup, meaning that they may not be charged and could go on to commit further crimes. Therefore, constructing lineups that allow witnesses to correctly distinguish between guilty and innocent suspects is paramount to ensure guilty individuals are identified and held responsible for their crimes, while innocent individuals are protected from being sanctioned for crimes that they did not commit.

The task of constructing fair lineups that allow for correct identification of the guilty suspect may appear relatively simple. Yet, a recent review of lineup guidelines of 54 countries world-wide highlighted the extent of variation in lineup construction and administration methods (Fitzgerald et al., 2021). Areas in which inconsistency was prevalent included witness instructions (i.e., whether they are told the perpetrator may not be present);

how fillers are selected (i.e., by match to the suspect's appearance or description) and their resulting similarity or dissimilarity to the suspect; lineup presentation (i.e., all lineup members presented at the same time, or individually); how many people are in the lineup; and how the lineup is presented, such as live, using photos or videos (Fitzgerald et al., 2021). Variation in lineup construction and administration methods in practice is likely to translate to variations in eyewitness identification accuracy.

Law enforcement and policy makers may look to the academic community for guidance on how best to construct and conduct lineups to maximise eyewitness accuracy. However, there is also a lack of consensus from academics on optimal lineup construction methods (e.g., Fitzgerald et al., 2021; Wells et al., 2020). For example, considering how the fillers should be chosen. Previously, researchers advocated for the construction of lineups using fillers who have been matched to the appearance of the suspect, in order to protect the innocent suspect from misidentification (Lindsay & Wells, 1980). Whereas others argue this could lead to selection of fillers that are highly similar, making it difficult for witnesses to correctly identify the guilty suspect if they are present in the lineup, due to the degree of overlap between fillers and the perpetrator in appearance (Luus & Wells, 1992). As such, most academics suggest the use of match to description strategies, in which lineup fillers are selected on the basis of their match to the witness's description of the perpetrator (Carlson et al., 2019; Luus & Wells, 1991; Navon, 1992; Wells et al., 1993; Wells et al., 1998). For example, if the witness describes the perpetrator as "white male, dark hair, small eyes", the lineup fillers should match on these features. Yet, the issue is further complicated by experimental design decisions, such as whether fillers are matched to the perpetrator (i.e., the guilty suspect) in both target absent and target present conditions, or to the guilty suspect in target present conditions, and to the innocent suspect in target absent conditions (Clark & Tunnicliff, 2001; see Chapter 2 for a full discussion of this issue). These experimental

decisions likely impact the similarity relationships between lineup members and therefore subsequent witness lineup performance (Quigley-McBride & Wells, 2021) and the conclusions that are drawn about optimal lineup construction methods. This issue is considered in more detail, next.

Filler Similarity

Within the present thesis, it is argued that the impact of filler selection methods can vary the similarity of fillers to both the guilty suspect and innocent suspect, and this is likely to impact the witness's ability to recognise the guilty suspect and reject an innocent suspect. As noted previously, in the academic community, there has been debate over the impact of selecting fillers that are highly similar or dissimilar to the suspect. For example, a previous meta-analysis suggested that high similarity lineups offered the best protection to innocent suspects from misidentification (Fitzgerald et al., 2013). However, a further review of the literature and an experiment highlighted the danger that highly similar fillers can reduce correct identifications of guilty suspects (Fitzgerald et al., 2015). The authors subsequently suggested selecting fillers of moderate similarity to the suspect was best practice (Fitzgerald et al., 2015). These examples highlight how academics have not yet been able to establish what optimal filler similarity looks like in practice (Fitzgerald et al., 2013, 2015; Wells et al., 2020). In chapter 2, a systematic literature review is conducted to identify optimal filler similarity conditions for subsequent lineup identification performance.

Innocent Suspect Selection

Moreover, the selection of the innocent suspect in experiments is not standardised within the literature and can alter conclusions. Some academics have argued that an innocent suspect, who is of median similarity to the guilty suspect, should be used in experiments because this represents the average range of the possible similarities of an innocent suspect in real-world police lineups (Colloff et al., 2021; Quigley-McBride & Wells, 2021). However

others have noted that in real life, a suspect may be selected because they resemble a footage from a crime scene (i.e., CCTV) or composite sketch of the perpetrator, and in these circumstances it is expected that the suspect will strongly resemble the perpetrator (Quigley-McBride & Wells, 2021; Wells & Penrod, 2011; Wixted & Wells, 2017). Furthermore, other research has found that a description matched, innocent suspect of moderate similarity to the perpetrator resulted in false identifications that were as many as the number of times the actual perpetrator was identified (Clark & Tunnicliff, 2001). Furthermore, it is reported that selecting an innocent suspect who is highly similar to the perpetrator creates a bias within the lineup so that there is an increase in identifications of the innocent suspect (Quigley-McBride & Wells, 2021). However, matching the innocent suspect and fillers to the description of the perpetrator has been argued to avoid a "backfire effect" in which the innocent suspect attracts a large number of false alarms (Clark & Tunnicliff, 2001; Wells et al., 1993). Furthermore, Quigley-McBride and Wells (2021) recommend "counterbalancing" in which a filler in the target absent lineup is replaced by the perpetrator in target present lineups (see Colloff et al., 2016). It is argued that this method enables the researcher to calculate the rate of innocent suspect identifications without bias (Quigley-McBride & Wells, 2021). As such, the variation within the literature on innocent suspect selection highlights the need for further analysis of the impact of methodological decisions on lineup outcomes. Therefore, the present thesis seeks to explore the impact of innocent suspect selection on subsequent filler-suspect similarity and lineup identification performance, as this can have implications for conclusions that are drawn for recommended practice (see chapter 3 for a full explanation).

Distinctive Suspects

Another area of concern is constructing lineups for suspects with distinctive facial features, such as scarring, a black eye or facial tattoos. It is well known that policing practice requires an identification procedure to be fair, meaning that the suspect should not stand out

from the other fillers (Police and Criminal Evidence Act, 1984, Code D, 2011). However, the task of ensuring a suspect with a distinctive facial feature does not stand out is not currently standardised within policing lineup methods. This is important because it is estimated that up to a third of lineup suspects have distinctive facial features (Flowe et al., 2018). In the UK, fair lineups may be constructed using concealment techniques, such as placing a block or pixelation over the distinctive feature of the suspect and fillers in the same area (see figure 5). Another method involves replication of the suspect's distinctive feature across the other lineup members (see figure 6). And both concealment and replication techniques are usually administered digitally. However, the implementation of the replication method is open to interpretation of the individual constructing the lineup. For example, a replication lineup for a distinctive suspect with a tattoo on the right cheek, could include fillers with the same tattoo on the right cheek, or the tattoo style and location could be varied across fillers (see figure 6). This thesis will consider why standardising replication methods is required to enhance witness identification performance in chapter 3.

Signal Detection Models

It is clear that there is currently a lack of academic consensus on lineup construction methods and their subsequent impact on lineup identification performance. The problem is potentially because the field has not been guided by a formal theory or model of memory, which has made it difficult to make concise predictions about the effect of similarity on lineup performance. As such, the present thesis considers the problem of lineup construction and witness identification performance through the lens of signal detection models; namely the feature matching model (Colloff et al., 2021) and diagnostic feature detection theory (Wixted & Mickes, 2014).

Essentially, the feature matching model (Colloff et al., 2021) assumes that the human face has a number of features (i.e., nose, eye, mouth) and each feature has a number of

settings (i.e., eye colour may be blue, brown, hazel, grey or green). It is this uniqueness of the perpetrator's features that can be used to enhance witness lineup identification (Colloff et al., 2021). Now, consider a lineup in which there is a suspect and fillers who have been matched to the witnesses' description of the suspect (i.e., brown hair, Caucasian, blue eyes etc). Within this lineup, some of the features of the perpetrator will be shared by other lineup members; these are the features that were in the witness's description (i.e., all lineup members will be Caucasian with blue eyes). However, the perpetrator will have a number of unique features, and their settings will have been observed by the witness and encoded in memory but not included in the witness's description. Because these features were not described by the witness, they will not be shared by all of the other members in a description-matched lineup. Therefore, these unique features of the perpetrator are diagnostic of guilt and can be relied upon by the witness to make their identification decision (Colloff et al., 2021). As such, lineup construction methods that enable the witness to focus on the unique features of the perpetrator, that are diagnostic of guilt, will result in the witness being more able to correctly identify the guilty suspect and correctly reject the innocent suspect.

Moreover, diagnostic feature detection theory (Wixted & Mickes, 2014) also states that witnesses will consider if features are shared across lineup members, and discount those that are shared, as they are no longer useful indicators of guilt. Features that are not shared across lineup members will be the focus of witness attention (and used to make the identification decision) as these features are more likely to be diagnostic of perpetrator guilt (Wixted & Mickes, 2014).

In the present thesis, the feature matching model (Colloff et al., 2021) and diagnostic feature detection theory (Wixted & Mickes, 2014) are used to predict optimal lineup construction methods for selecting fillers (chapter 2) and for distinctive suspects (chapter 3) to enhance eyewitness identification performance. To provide an overview, it follows that

conditions of high similarity (including highly similar fillers or replication of the same distinctive facial feature) would lead to a higher degree of overlap (i.e., more shared features) between the memory signals for the guilty suspect and fillers in witness memory. And there will be fewer diagnostic features for the witness to be able to correctly identify the guilty suspect and reject the innocent suspect, resulting in a decreased identification of the guilty suspect (Colloff et al., 2021). Alternatively, conditions of lower similarity that are still fair because the suspect does not stand out, should result in less overlap (i.e., less shared features) between the guilty suspect and fillers. And there will be more available diagnostic features that the witness can use to correctly identify the guilty suspect when they are presented (Colloff et al., 2021). In target absent conditions, lower similarity lineups prevent the innocent suspect from misidentification, as the innocent suspect does not have a stronger memory signal than the other fillers in the lineup (Colloff et al., 2021). Therefore, in target absent conditions, manipulating filler similarity does not impact innocent suspect identifications, and this is because the innocent suspect is no more similar to the guilty suspect than the other lineup members. Moreover, any features that are shared by both the innocent suspect and guilty suspect will match by chance.

Discriminability

Witness's collective ability to tell the difference between guilty perpetrators and innocent suspects is known as "discriminability". Empirical discriminability refers to the degree to which witnesses are able to accurately sort innocent and guilty suspects into their respective groups (Wixted & Mickes, 2018). Whereas theoretical discriminability refers to the amount of theoretical overlap between the memory strengths for innocent and guilty suspects in the witness's memory (Wixted & Mickes, 2018). It has been argued that to improve lineup performance, researchers should seek to construct lineups that enhance discriminability (NRC, 2014). That is, increased identification of the guilty suspect (known

as the Hit Rate, HR) and decreased misidentification of the innocent suspect (known as the False Alarm Rate, FAR). The HR is calculated by dividing the number of times the guilty suspect was identified, by the number of target present lineups administered. The FAR is calculated by dividing the number of times the innocent suspect was identified by the number of target absent lineups presented. To calculate witness's ability to discriminate between innocent and guilty lineup members, discriminability analysis can be conducted using the conceptual formula provided by Mickes et al. (2014); d' = z(correct ID rate) – z(false ID rate). This results in a discriminability measure known as d-prime (d'). A higher d' value indicates a better ability for the witness to discriminate between the innocent and guilty suspect and a d' value of 0 indicates an inability of witnesses to discriminate between innocent and guilty suspects (Macmillan & Creelman, 2004).

Measurement of discriminability is of particular interest due to academic debate on how best to measure eyewitness performance. Some research has concluded that fair lineups enhance discriminability compared to unfair lineups in which the suspect stands out (Colloff et al., 2016, 2017). And it has been argued that these findings can be explained by the mechanisms of diagnostic feature detection theory (Colloff et al., 2016, 2017; Wixted & Mickes, 2014). That is, fair lineups are advantageous over unfair lineups, because the witness is able to discount features that are non-diagnostic of guilt, i.e., facial features shared by lineup members. Whereas in unfair lineups, in which the distinctive suspect stands out, the witness may focus on the distinctive feature, even though it is non-diagnostic of guilt in target absent conditions and therefore discriminability is impaired. However, an alternative filler siphoning perspective (Smith et al., 2019, 2022) argues that fair lineups do not enhance ability to discriminate between innocent and guilty suspects. Instead, it is suggested that a fair lineup advantage exists because fair lineups lead to a distribution of choices away from the suspect and onto fillers, and this is known as 'filler siphoning' (Smith et al., 2018). While it is

acknowledged that filler siphoning is possible and certainly occurs (i.e., more fillers are chosen in fair lineups), it has been argued that the filler siphoning alone cannot explain why fair lineups improve ability to discriminate innocent from guilty suspects, because filler siphoning does not make a prediction about theoretical discriminability (Colloff et al., 2018). However, the filler siphoning perspective does predict that in fair lineups, there will be improved ability of the investigator to discriminate innocent from guilty suspects (Smith et al., 2020; Smith et al., 2022). That is, in fair lineups, where the investigator knows which lineup members are fillers, there will be an improved ability of the investigator to use eyewitness evidence to discriminate between innocent and guilty suspects (Smith et al., 2020; Smith et al., 2022). And as the debate continues, it is clear that further investigation is required to establish which lineup construction methods result in enhanced discriminability and further test signal-detection based model explanations of the underlying memory processes involved in witness identification decisions (e.g., Colloff et al., 2021; Wixted & Mickes, 2014).

Individual Differences

So far, this thesis has considered the impact of lineup construction methods on resulting witness discriminability in a global manner. That is, it has been assumed that any change in discriminability is due to the lineup condition alone and considered which lineup condition, collectively over all witnesses, enhances performance. However, a limitation of this approach, used both in the present research and within the lineup literature as a whole, is the failure to consider the potentially confounding impact of individual differences in witness ability to recognize the perpetrator. Research has highlighted than within the general population, some people may be better at facial recognition than other people (Wilmer et al., 2010; Zhu et al., 2010). Moreover, experiments have identified a correlation of eyewitness accuracy and performance on facial recognition tests (Binderman et al., 2012; Geiselman et

al., 2003; Hosch, 1994; Memon et al., 2003; Morgan et al., 2007; Searcy et al., 1999; Searcy et al., 2001). Therefore, the use of facial recognition tests may be beneficial to evaluate the impact of the witness's facial recognition ability on subsequent lineup identification performance. This could be applied to research to establish if lineup manipulations enhance performance over the range of individual differences in facial recognition abilities. And in police practice, a measure of witness facial recognition ability could be used to evaluate the likely accuracy of a witnesses' identification decision. In chapter 4, the feasibility of the Benton Facial Recognition Test for assessing individual differences in facial recognition within the eyewitness arena will be explored.

Thesis Aims and Outline

The present thesis aims to:

1. Investigate optimal lineup construction methods that enhance discriminability, specifically considering the impact of suspect filler similarity and distinctive facial features
2. Test the feature matching model and diagnostic feature detection theory accounts of eyewitness identification decision making
3. Evaluate the use of the Benton Facial Recognition Test and its applicability to considering individual differences in witness identification performance.

Thesis Outline

Chapter 1 has provided an overview of witness lineup identification through the lens of models based in signal detection theory and the need for lineup construction methods that enhance discriminability (see Appendix G for a glossary of terminology used in this thesis).

In chapter 2, the existing suspect-filler similarity literature is re-examined through the lens of signal detection theory. Predictions are made about optimal lineup construction on eyewitness discriminability using feature matching model (Colloff et al., 2021) and

diagnostic feature detection theory (Wixted & Mickes, 2014). A systematic literature review was conducted and identified 29 experiments to be included in discriminability analysis. This review included a calculation of d' (conceptual formula by Mickes et al., 2014) for the included experiments and tested feature matching model predictions (Colloff et al., 2021) to observe if the model-based hypotheses were borne out in the existing literature. In doing so, the impact of line up construction methods in experiments (i.e., filler and innocent suspect selection) on suspect-filler similarity dynamics and overall discriminability of lineup conditions, is considered to help try to explain conflicting results in the field to date.

In chapter 3, an experiment of lineup construction for distinctive suspects was conducted. In which, there is a comparison of a high similarity replication lineup, in which the distinctive suspects' facial tattoo is exactly replicated across lineup members; a low similarity replication lineup, in which the lineup members have a similar but non-identical distinctive facial tattoo; and an unfair do-nothing lineup in which the distinctive suspect is the only lineup member with a facial tattoo. Then, the experiment considered empirical discriminability for each lineup condition using area under the ROC curve (AUC) statistical analysis.

In chapter 4, the applicability of Benton Facial Recognition Test (BFRT) to assessing individual differences in facial recognition was considered. In which the evidence base for the BFRT is detailed, including several versions of the BFRT. And this chapter explores whether the BFRT accurately measures facial recognition ability and identifies when there are deficits in facial processing abilities.

Finally, chapter 5 presents a review mixed findings within the literature to date. Crucially, it was concluded that low similarity conditions are optimal for increasing discriminability and do not increase the risk of misidentification when the fillers and innocent suspect are from the same memory distribution. It was argued that this is because low

similarity lineups allow the witness to focus on the perpetrators' unique features that are diagnostic of guilt, to make an accurate identification decision (Colloff et al., 2021: Wixted & Mickes, 2014). Furthermore, chapter 5 highlights that the methodological variances both within the reviewed literature of chapter 2 and within policing practice worldwide (Fitzgerald et al., 2021) explain why it has not been possible to identify optimal lineup construction methods that consistently increase discriminability across lineup conditions.

CHAPTER 2: A SYSTEMATIC REVIEW OF LINEUP FILLER SIMIALRITY

Abstract

Within eyewitness research it is agreed that fair lineups, in which the suspect does not stand out from other lineup members, should be constructed. However, there is currently a lack of academic consensus on the optimal level of filler similarity that best supports eyewitness identification accuracy. The present review sought to explore optimal filler similarity by examining relevant literature through the lens of signal detection theory. Experiments were identified through a systematic search of electronic databases, article reference lists and contacting key experts. Following application of inclusion/exclusion criteria, quality assessment and initial data extraction, twenty nine experiments published between 1980 and 2022 were included within the present review. Diagnostic feature detection theory and feature matching model (Colloff et al., 2021; Wixted & Mickes, 2014) were used to make predictions about how filler similarity conditions within existing research should influence perpetrator identifications (hit rate), innocent suspect identifications (false alarm rate) and witness ability to discriminate between innocent and guilty lineup members (d'). A review of trends within the literature indicated that the predictions of signal detection theories were mostly supported within the literature and suggest that signal-detection based theories offer valuable insight into the processes involved in witness identification. It was highlighted how methodological characteristics can influence similarity comparisons and lineup identification outcomes. It was recommended that future research reports lineup construction methodology and provides open access to experiment materials for replication and future reviews. Finally, low similarity lineup conditions were suggested as the optimal approach; however this was conditional on methodological characteristics, such as ensuring the lineup is fair so that the suspect does not stand out.

Introduction

Police can use witness memory to help apprehend the perpetrator of a crime. A witness may be asked to identify the perpetrator from a lineup identification procedure. A lineup typically includes a suspect (who may be the perpetrator or an innocent suspect) and other individuals (referred to as "fillers") who are similar in physical resemblance to the suspect but are known to be innocent. Within the literature, a distinction is made between lineups that contain the perpetrator (i.e., the guilty suspect), known as "target present" and lineups that contain an innocent suspect called "target absent". Furthermore, lineups have been constructed to contain fillers to protect innocent suspects. This is critical when the witness is inclined to make an identification in the lineup, even when they are not certain of their own identification accuracy. Protection of the innocent suspect is highly important to ensure that innocent individuals are not wrongly convicted, and guilty perpetrators are appended for their crimes (NRC, 2014; Quigley-McBride & Wells, 2021).

Academics and policymakers agree that the most appropriate way to protect an innocent suspect is to construct fair lineups whereby the suspect does not stand out (National Institute of Justice, 1999; Police and Criminal Evidence Act, 1984, Code D, 2011; Wells et al., 2020). This was supported by a recent review of the eyewitness literature, which recommended that lineups should contain "at least five appropriate fillers who do not make the suspect stand out in the lineup based on physical appearances of other contextual factors such as clothing or background" (Wells et al., 2020, p.8). Research has found that unfair lineups also impair the witness's ability to differentiate between innocent and guilty lineup members compared to fair lineups (Clark, 2012; Colloff et al., 2016, 2017; Wells et al., 1979). Others argue that the advantage of fair lineups exists because they lead to a distribution of choices away from the suspect and onto fillers, resulting in improved discriminability of the outside observer, who is aware of which lineup members are fillers

(Smith et al., 2018; Smith et al., 2022). Despite ongoing theoretical debate, it is agreed that fair lineups should be utilised in practice. Nevertheless, it is still not clear which method of constructing a fair lineup best optimises witness identification accuracy (Wells et al., 2020).

How are lineups constructed in practice?

Firstly, the "match to appearance" strategy is when fillers are selected based on physical similarity to the suspect (Luus & Wells, 1991; Wogalter et al., 2004). Secondly, the "match to description strategy" is when fillers are selected based on the description of the perpetrator provided by witnesses or on default variables such as facial hair, age, and gender in the absence of a suitable description (Lindsay et al., 1994; Luus & Wells, 1991; Wells et al., 1993). In a review of eyewitness identification guidelines in 54 countries, it was reported that 89% of guidelines recommend using the match to appearance strategy, whereas only 17% of guidelines endorsed the match to description method (Fitzgerald et al., 2021). Where the match to appearance method is recommended, there is typically no mention of how to consider the witness's description, except for guidelines in Scotland which state that fillers matching the suspect appearance is more important than matching the witness description (Fitzgerald et al., 2021; Police Scotland, 2018). Additionally, in the guidelines that endorse the match to description method, five countries also recommend use of match to appearance (Fitzgerald et al., 2021). That is, the fillers should both match the witness's description and also match the appearance of the suspect.

The use of match to appearance or description methods to construct lineups is of significance because research shows that the different lineup selection strategies result in lineups that contain fillers that differ in similarity (Quigley-McBride & Wells, 2021). The match to appearance strategy could become problematic if fillers selected are highly similar to the perpetrator, as this could make it more difficult for the witness to correctly identify the perpetrator due to the degree of overlap between fillers and perpetrator in physical

appearance (Luus & Wells, 1991; Wells et al., 1993)). In contrast, the match to description strategy may be less likely to result in highly similar fillers as the description provides the parameter on which filler similarity can be based (Colloff et al., 2021; Luus & Wells, 1991; Wells et al., 1993). Accordingly, some research has found that compared to match to description methods, the match to appearance strategy can result in fewer perpetrator identifications (Juslin et al., 1996; Wells et al., 1993). For this reason, some research has advocated for the use of match to description strategies (Carlson et al., 2019; Luus & Wells, 1991; Navon, 1992; Wells et al., 1993; Wells et al., 1998). Furthermore, Wells et al. (1993) compared description matched fillers that were either highly similar or highly dissimilar to the suspect and found that the number of identifications of the guilty suspect was lower when highly similar fillers were presented in the lineup. And Wells et al. (1993) suggested that using fillers that match the witness's description of the perpetrator, but that do not resemble each other (i.e., low similarity) would result in increased identification of the guilty suspect while protecting the innocent suspect. Other research, however, has reported no difference in eyewitness performance on lineups in which fillers have been matched to appearance versus matched to description (e.g., Darling et al., 2008; Lindsay et al., 1994; Tunnicliff & Clark, 2000).

An alternative approach is to use a combination of match to appearance and description methods (Fitzgerald et al., 2021; Lindsay et al., 2007). This involves using the witness's description to create a pool of plausible fillers who are matched to the suspect description, and then selecting fillers from that pool, who are similar to the suspect, to be used in the lineup (Fitzgerald et al., 2021; Lindsay et al., 2007). The use of this approach seeks to produce lineups that are "as fair as possible" by protecting the innocent suspect. However, a combination approach also makes the identification task harder due to increased similarity between the suspect and fillers (Wells et al., 1993). To overcome this problem,

Colloff et al. (2021, p.5) recommended the following: "from a pool of acceptable description-matched photos, select fillers who are dissimilar to the suspect". This recommendation, consistent with the findings of Wells et al. (1993), was supported by the finding that the use of dissimilar fillers increased the correct identification of the guilty suspect and did not affect the identification rate of the innocent suspect (Colloff et al., 2021). To date however, there is not a clear picture from the literature to advise police practice on lineup construction and it is acknowledged within the field that the issue of lineup construction is not resolved (Wells et al., 2020).

Existing Reviews on Filler Similarity

Despite the differences in methodological approaches and findings drawn across experiments, some researchers have attempted to review the literature to determine which lineup methods should be recommended in practice because they increase the identification of perpetrators and decrease the identification of innocent suspects. In a meta-analytic review, Fitzgerald et al. (2013) identified eleven experiments that manipulated filler similarity. Experiments included in the review utilised a mock crime paradigm whereby participants viewed a staged crime and were required to identify the perpetrator from a lineup which contained which low, moderate and/or high-similarity fillers.

This review highlighted that low similarity lineups resulted in more suspect (guilty and innocent) identifications than moderate and high similarity lineups. Additionally, Fitzgerald and colleagues found that moderate and high similarity lineups resulted in more filler identifications than low similarity lineups. It was also reported that in target present lineups containing the perpetrator, high similarity fillers were harmful to witness identification accuracy, reducing the hit rate (Fitzgerald et al., 2013). Conversely, in target absent lineups containing an innocent suspect, high similarity fillers were beneficial to witness identification accuracy, reducing the false alarm rate (Fitzgerald et al., 2013). The

authors advocated the use of high similarity lineups to offer the best protection to innocent suspects from misidentification (Fitzgerald et al., 2013). However, after a further review of the literature and an experiment, Fitzgerald et al., (2015) highlighted the potential danger of very highly similar fillers on reducing correct identifications and suggested moderate similarity fillers should be used as best practice.

The mixed conclusions of existing experiments and reviews highlight how academics have not been able to establish what filler similarity should be employed in practice to optimise witness identification accuracy (Fitzgerald et al., 2013, 2015; Wells et al., 2020). It is possible that the existing literature and reviews have been unable to establish optimal filler similarity as the impact of experiments' methodological characteristics on filler similarity manipulations (and therefore the experimental results) have not been considered. In previous reviews (Fitzgerald et al., 2013, 2015) findings of all experiments were considered together, regardless of methodological inconsistencies across studies. For example, similarity was categorised based on the degree of resemblance of the fillers to the suspect (both guilty and/or innocent). Additionally, the problem is potentially because the field has not been guided by a formal theory or model of memory, which has made it difficult to make concise predictions about the effect of similarity on lineup performance.

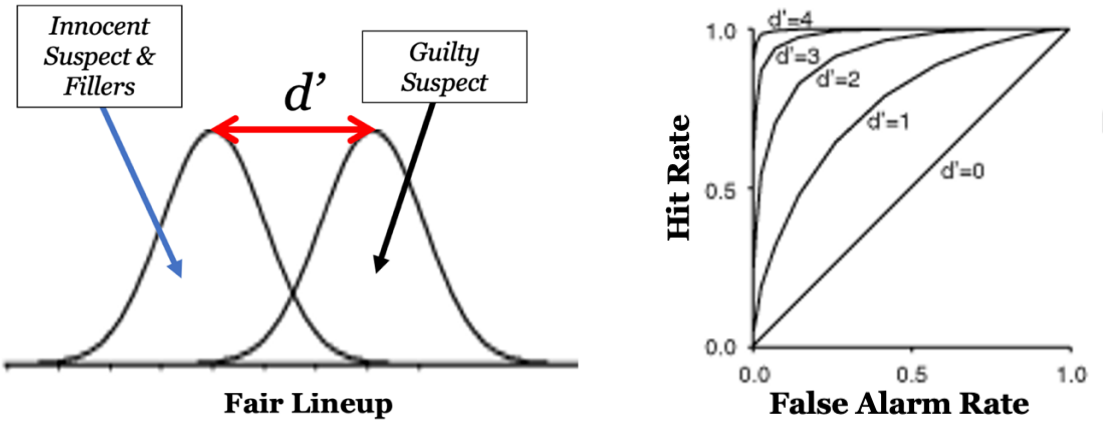
Signal Detection Theory

Considering eyewitness identification through the lens of signal detection theory should help us to better understand the optimal filler similarity in lineups (Wixted & Mickes, 2014). According to a signal detection interpretation of a lineup task, each lineup member generates a memory signal within the witness. Across witnesses and lineups, these memory signals can be displayed as three memory strength distributions: one for perpetrators (guilty suspects), one for innocent suspects, and one for fillers. When a lineup is fair, the innocent suspect and filler distributions are the same and will therefore overlap, leading to a two

distribution model of suspects and fillers/innocent suspects (see figure 1a). The degree of overlap between these memory distributions is known as discriminability, with more overlap indicating poorer discriminability (i.e., poorer ability to tell the difference between lineup members; Wixted & Mickes, 2014). When the variances of the memory strength distributions are equal, the distances can be measured by d' (Mickes et al., 2014), known as a measure of theoretical discriminability. Empirical discriminability (i.e., requiring no assumptions about underlying memory strength distributions) can be measured by Receiver Operating Characteristic (ROC) analysis. An ROC plot depicts the hit rate (perpetrator identifications) and false alarm rate (innocent suspect identifications). Lineup conditions yielding higher discriminability result in an increased hit rate of correct perpetrator identifications, and a decreased false alarm rate of innocent suspect identifications and are depicted in a higher ROC curve as shown in figure 1b. As also evident in figure 1b, higher ROC curves (an empirical measure of discriminability) typically equate to larger d' values (a measure of theoretical discriminability; see Mickes et al., 2014).

Figure 1

Two distribution model of memory for innocent suspect/fillers and guilty suspects in a fair lineup (1a) and corresponding ROC curves for lineup conditions (1b). d' measures discriminability and represents the overlap in memory signals for innocent suspect and fillers, and the guilty suspect. The larger d' is, the smaller the overlap of the distributions in memory.



1a) **1b)**

Signal Detection-Based Models of Eyewitness Memory

To consider lineup conditions that yield higher discriminability, two models based in signal detection theory have been proposed in the lineup literature: diagnostic feature detection theory (Wixted & Mickes, 2014) and the feature matching model (Colloff et al., 2021). These models can help us to make predictions about optimal filler similarity conditions.

Feature Matching Model (Colloff et al., 2021)

This model assumes that a face is defined by a number of features and that each facial feature has several possible settings (Colloff et al., 2021). For example, the feature of eye colour may have settings of brown, blue, hazel, grey and green. And, after witnessing a crime, the witness will have stored in memory the unique features of the perpetrator's face. When presented with a lineup in which the perpetrator is present, the encoded features of the

perpetrator in the witnesses' memory will match those of the perpetrator presented in the lineup. However, an innocent suspect and fillers in a lineup, who are not guilty, will not possess the same matching features as they are unique to the perpetrator. In a description matched lineup, fillers are selected for the lineup on the basis that they match the witness description, and so some of the perpetrator's features will be shared by the fillers and innocent suspect, and these features will be non-diagnostic of guilt. However, the perpetrator will possess unique features that are not shared by the fillers or innocent suspect in the lineup (i.e., those not in their description), which are diagnostic of guilt and can be relied upon by the witness in making an identification decision. Therefore, lineup conditions which maximise the ability of the witness to focus on facial features that are diagnostic of guilt to make an identification decision, will improve witness accuracy.

It is possible to make predictions about optimal filler similarity in fair lineups using the feature matching model. Note that, in all cases considered next, all fillers match the witness's description of the perpetrator. In *target present conditions*, higher similarity fillers will share many features that match the witness's memory of the perpetrator, reducing the number of unique features on the perpetrator in the lineup that match the witness's memory of the perpetrator. As such, high similarity fillers compete with the witness's memory of the perpetrator, resulting in a decrease in the hit rate. Lower similarity fillers will share fewer features that match the witness's memory of the perpetrator, increasing the number of unique features on the perpetrator that match the witness's memory of the perpetrator. As such, lower similarity fillers compete with the witness's memory of the perpetrator to a lesser extent, resulting in an increase in the hit rate.

The predictions of the model differ depending on how target-absent lineups have been constructed. In target absent conditions, where filler similarity to the innocent suspect has been manipulated (i.e., suspect matched lineups), varying filler similarity should not change

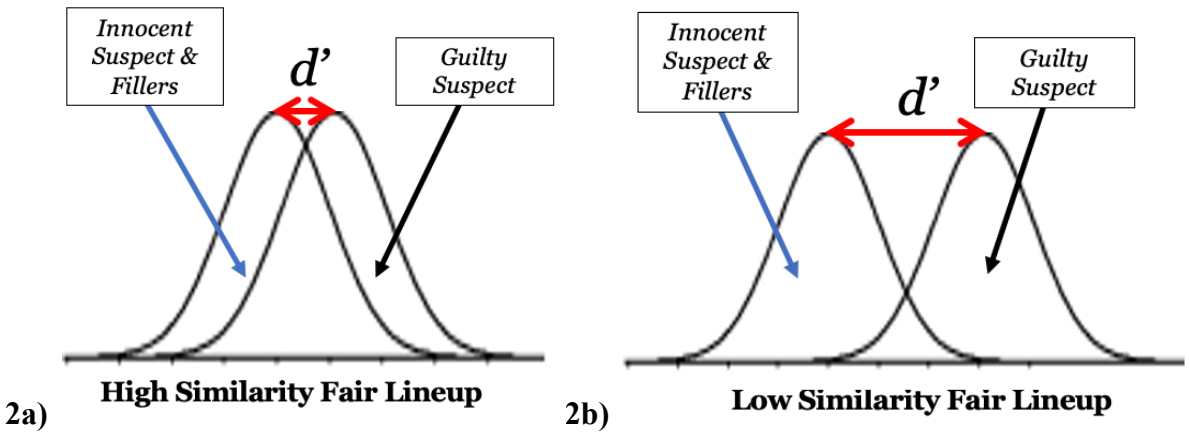
the number of features on the face of the innocent suspect that match the witness's memory of the perpetrator. Therefore, the false alarm rate should remain unchanged across lower and higher filler similarity conditions.

Diagnostic Feature Detection Theory (Wixted & Mickes, 2014)

This theory argues that when making a lineup identification decision, optimal witnesses will discount shared lineup member features that are non-diagnostic (i.e., features that are not indicative of guilt because they are shared across all lineup members) and instead focus on diagnostic features that are not shared across all lineup members (Wixted & Mickes, 2014). It is also possible to make predictions about optimal filler similarity in fair lineups using diagnostic feature detection theory. In *high similarity lineups* the availability of potential diagnostic features is reduced due to the resemblance between the fillers and perpetrator, reducing discriminability. Whereas in *low similarity lineups* all lineup members will share some non-diagnostic features that have been used to match the fillers to the perpetrator or innocent suspect, however there are still diagnostic features available to make an identification decision, increasing discriminability. This is depicted in figure 2 below.

Figure 2

Two distribution model of memory for innocent suspect/fillers and guilty suspects in a high similarity fair lineup (2a) and a low similarity fair lineup (2b).



2a) **2b)**

Critically, however, the pattern of results predicted by the feature matching model and diagnostic feature detection theory, is impacted by the methodological decisions made in experiments and the characteristics of the lineup. When considered using a signal detection theory framework, (i.e., considering the problem in terms of memory strength distributions), it becomes clear why the pattern of results found in previous experiments is mixed. Namely, because of methodological differences across experiments change how lineup member similarity impacts on eyewitness identification. Some methodological decisions result in the innocent suspect in target absent lineups becoming more similar to witness's memory of the perpetrator than the other lineup members, which would increase the false alarm rate to the innocent suspect in low similarity lineups and impair discriminability. Those methodological decisions are described next.

Suspect or Perpetrator Matched?

Experiments vary in how the experimental target absent conditions are created, which may influence identification accuracy and experiment conclusions (Colloff et al., 2021; Oriet & Fitzgerald, 2018). In "perpetrator matched" experiments, fillers are selected based on their match to the guilty perpetrator in both target present and target absent lineup conditions (Clark & Tunnicliff, 2001). That is, low similarity target present conditions would contain fillers that are low similarity to the perpetrator. And in target absent conditions, they would contain fillers that are low similarity to the perpetrator but would not contain the perpetrator and would instead have an innocent suspect. In "suspect matched" experiments, fillers are matched to the guilty perpetrator in target present conditions and the innocent suspect in target absent conditions (Clark & Tunnicliff, 2001). That is, low similarity target present conditions would contain fillers that are low similarity to the perpetrator. And in target absent conditions, they would contain fillers that are low similarity to the innocent suspect.

In target absent conditions, the two strategies—perpetrator matched or suspect matched—appear to influence identification outcomes. Research has highlighted that in target absent perpetrator matched lineup conditions, where filler similarity has been manipulated relative to the perpetrator, the innocent suspect stands out because the fillers are matched to the guilty perpetrator and not the innocent suspect (Clark & Tunnicliff, 2001). In low similarity conditions, this results in an increase in false alarms to the innocent suspect (Colloff et al., 2021). However, in suspect matched lineups, where filler similarity has been manipulated relative to the innocent suspect, research found no difference in the false alarm rate of the innocent suspect across conditions that vary in filler-suspect similarity. Put another way, low similarity fair lineups, on average, do not put the innocent suspect at increased risk of being falsely identified (Colloff et al., 2021; Oriet & Fitzgerald, 2018).

The feature matching model and diagnostic feature detection theory predictions outlined earlier apply to the suspect matched approach. Importantly, the opposite pattern of results is predicted when the perpetrator matched approach is used. The feature matching model predicts that, in *target absent conditions* where the filler similarity to the perpetrator has been manipulated (i.e., perpetrator matched lineups), varying filler similarity will cause the innocent suspect to stand out in memory in lower similarity conditions, because the innocent suspect shares more features with the perpetrator in memory than do the other fillers, and therefore the false alarm rate will increase. Moreover, fewer shared features can be discounted when lower similarity compared to higher similarity fillers are used, reducing discriminability in low similarity compared to high similarity lineups (Colloff et al., 2021; Wixted & Mickes, 2014).

Innocent Suspect Selection

Experiments also vary in how the innocent suspect is selected. It was suggested by Colloff and colleagues that, from a pool of faces, an innocent suspect that is of median

similarity to the perpetrator should be selected because this is more representative of the range of possible innocent suspects that could be selected in real life lineups (Colloff et al., 2021). However in real life, a suspect may be selected because they resemble a footage from a crime scene (i.e., CCTV) or composite sketch of the perpetrator, and in these circumstances it is expected that the suspect will strongly resemble the perpetrator (Quigley-McBride & Wells, 2021; Wells & Penrod, 2011; Wixted & Wells, 2017). Moreover, other researchers have argued that the use of a moderate similarity description matched innocent suspect resulted an increased false alarm rate, which they called a "backfire effect" (Clark & Tunnicliff, 2001). To avoid this, Clark and Tunnicliff (2001) suggest the use of match to description, in which both the innocent suspect and fillers are matched to description of the perpetrator. Moreover, selecting the innocent suspect by match to description has also previously been recommended (Wells et al., 1993). On the other hand, some researchers use an innocent suspect who is highly similar to the perpetrator to compare the impact on subsequent lineup performance (e.g., Carlson et al., 2019; Gronlund et al., 2009). Furthermore, Quigley-McBride and Wells (2021) recommend "counterbalancing" in which a filler in the target absent lineup is replaced by the perpetrator in target present lineups. It is argued that this method enables the researcher to calculate the rate of innocent suspect identifications without bias (Quigley-McBride & Wells, 2021).

Clearly, there is variability across experimental research in how the innocent suspect is selected. And this is problematic because an innocent suspect who is highly similar to the perpetrator presented with moderate similarity fillers (or low similarity fillers), results in the innocent suspect being more likely to be wrongly identified by the witness due to being the best match to the perpetrator (Clark & Tunnicliff, 2001; Quigley-McBride & Wells, 2021). From a signal-detection framework, using an innocent suspect who is more similar to the witness's memory of the perpetrator than the fillers, on average theoretically results in three

memory distributions: one for the perpetrator, one for the innocent suspect who shares a higher proportion of the perpetrators features than the fillers, and one distribution for the fillers who have less of the perpetrator's unique features. Essentially, presenting an innocent suspect who is highly similar to the perpetrator along with moderate similarity (or low similarity) fillers, results in an unfair lineup. And the feature matching model and diagnostic feature detection theory (Colloff et al., 2021; Wixted & Mickes, 2014) predict the opposite pattern of results than outlined above when the lineup is unfair, which is explained next.

Low Similarity vs. Unfair Lineups

Within the literature, there has been limited consideration of the difference between unfair lineups, in which the suspect is the only plausible lineup member, and those in which fillers are of low similarity to the suspect but are still within the constraints of the witness description for the perpetrator (i.e., low similarity but fair lineups). Indeed, the Fitzgerald et al. (2013) meta-analysis found more guilty and innocent suspect identifications in low similarity lineups, suggesting that both the guilty and innocent suspect matched memory for the perpetrator more than the other fillers. Put another way, it is possible that the low similarity lineups in which innocent suspects were more regularly identified, were unfair.

When lineup fillers become so low similarity that they do not match the description provided by the witness (i.e., are unfair), the feature matching model and diagnostic feature detection theory predict the opposite pattern of results than outlined above. Namely, according to the feature matching model, in *target absent conditions* where unfair lineup fillers are used, varying filler similarity will cause the innocent suspect to stand out in memory to a greater extent in lower similarity conditions, because the innocent suspect shares more features with the perpetrator in memory than do the other fillers, and therefore the false alarm rate will increase. Moreover, the diagnostic feature detection theory (Wixted & Mickes, 2014) predicts that discriminability will be impaired in unfair compared to fair

lineups. In *fair lineups,* when the innocent suspect does not resemble the perpetrator any more than other fillers, the witness theoretically discounts non-diagnostic shared features and instead focuses on diagnostic features to make an identification decision, increasing discriminability. However, in *unfair lineups,* when the innocent suspect resembles the perpetrator more than other fillers, it is not clear that certain features on the face of the innocent suspect are non-diagnostic of guilt. As such, the witness is more likely to focus on non-diagnostic features to make an identification decision, decreasing discriminability. This predicted discriminability effect between fair versus unfair lineups has been observed across a number of experiments, though note that the differences have not always found to be statistically significant (Colloff et al., 2016, 2017; Wetmore et al., 2015; Key at al., 2017; Lucas et al., 2021; Flowe et al., 2021).

Present Review

There is no academic consensus on the optimal filler similarity to create lineups that protect the innocent and increase identification of the guilty (Wells et al., 2020). Experiments that have examined lineup member similarity vary methodologically, such as how the innocent suspect is selected, if fillers have been suspect or perpetrator matched, and whether match to appearance, description or a combination has been used (Quigley-McBride & Wells, 2021). Signal detection theory indicates that these methodological variations are likely to be important determinants for how filler similarity manipulations (e.g., high, medium, low similarity) influence eyewitness identification outcomes. However, methodological variations have not previously been explored in reviews of this area when interpreting the experiment outcomes (Fitzgerald et al., 2013, 2015). It is argued that methodological variations are important determinants of how filler similarity manipulations influence eyewitness identification outcomes (Colloff et al., 2021; Oriet & Fitzgerald, 2018; Quigley-McBride & Wells, 2021; Wixted & Mickes, 2014). Therefore, the present review seeks to reconsider

lineup filler similarity through the lens of signal detection theory, by considering the methodological decisions made in experiments and how these may have influenced the study conclusions.

When there are methodological limitations in experiments, systematic literature reviews have been recommended to identify gaps in the existing research (Cheung & Vijayakumar, 2016; Garg et al., 2008). A meta-analysis of the current literature was not deemed suitable for our current task of exploring methodological differences across studies. Because of the abovementioned methodological inconsistencies across experiments that are predicted to result in different patterns of results, it would not be possible to compute a single meta-analysis of the effect of lineup similarity of eyewitness outcomes. Instead, a systematic review is conducted to examine the trends within the data and test theoretical predictions about the likely pattern of results in each study, as predicted by signal detection models; feature matching model (Colloff et al., 2021) and diagnostic feature detection theory (Wixted & Mickes, 2014).

Therefore, the aims of this systematic review are to explore:

- How methodological characteristics of experiments influence experiment outcomes
- How lineup filler similarity impacts identification of the perpetrator (hit rate) and identification of the innocent suspect (false alarm rate)
- How lineup filler similarity affects witness discriminability

Method

Sources of literature

A scoping search of the Cochrane Library, Campbell Library and electronic databases was conducted to identify current reviews of similarity manipulations in eyewitness identification. Two existing papers were identified; a meta-analysis by Fitzgerald et al. (2013) and a review by Fitzgerald et al. (2015). Both articles were evaluated to assess if the

need remained for a further review in this area. It was established that previous reviews (Fitzgerald et al., 2013, 2015) supported the continued development of the field as they identified the dangers of highly similar fillers on reducing correct identifications. However, rationale for the present review remained as the impact of methodological inconsistencies within the literature on experimental outcomes has not previously been explored in a systematic review context.

Search strategy

The following hierarchy of search terms was used to identify relevant literature:

1	Witness* near/2 (identif* or accura* or confiden* or discriminability* or bias*)
2	Lineup near/3 (filler* or Foil* or Similar* or select* or match* or appear* or construct* or compos* or fair* or unfair*)
3	1 AND 2

Literature searches were completed between June 2020 and July 2020 in Scopus, EBSCO host, Web of Science, PubMed, ProQuest, PsychInfo. Additional articles published after the initial search were identified by database searching and contact with experts within this research area. Two identified articles were quality assessed and included in this review (Colloff et al., 2021; Lucas et al., 2020). Articles were identified through additional sources by reviewing reference lists of previous reviews (Fitzgerald et al., 2013, 2015). A second search of the same data bases and search terms was completed in April 2022. Two articles that were published since the previous search were identified and included in the present review (Lucas & Brewer, 2021; Smith et al., 2022). (See Appendix A for details of search record).

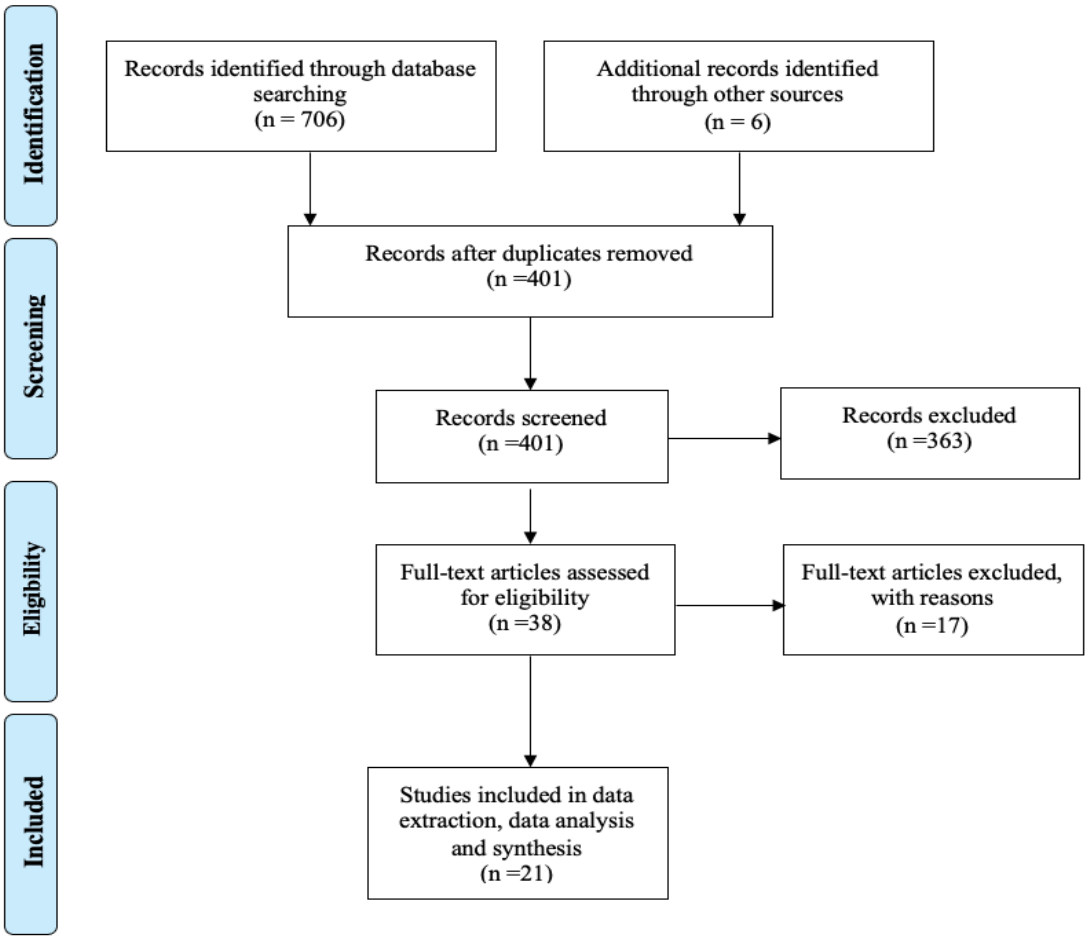
Article selection

Articles were screened to check they met inclusion criteria, those that failed to meet the criteria were excluded. Next, articles were assessed for methodological quality. Articles

were excluded if they were assessed as being poor in quality, outside of the scope of this review, or did not report sufficient outcome data for analysis (as it was not possible to calculate ability to discriminate between innocent and guilty lineup members).

Figure 3

PRISMA flow diagram of article selection process



Inclusion Criteria

Population. The review included experiments that required participants to encode a face (or set of faces) and later tested participant memory for the face(s) using a lineup identification procedure. Most experiments used a mock crime paradigm whereby participants witnessed a staged crime and then completed a lineup to identify the perpetrator from the mock crime event. Participants were adults aged 16 and over who had taken part in the experiments.

Intervention. Included experiments were required to test witness memory using a lineup task, which contained more than one person. Experiments using showups of one person were excluded. A lineup could be presented in simultaneous or sequential format. Experiments were included if they manipulated the similarity of the physical appearance of lineup members (perpetrator, innocent suspect, fillers). Experiments that did not manipulate physical appearance similarity (e.g., manipulated expression, hair colour, clothing) were excluded from the review. Research that also manipulated other variables (e.g., encoding conditions or lineup instructions) was only included if the other manipulated variables were fully crossed with the similarity manipulation. Experiments that manipulated variables within similarity conditions such as adding low similarity fillers to a lineup of high similarity fillers (see Charman et al., 2011; Nosworthy, & Lindsay, 1990) were excluded on the basis that the presence of other variables known to influence lineup identification would be confounded with the similarity manipulation and make it difficult to interpret the effect of the similarity manipulation alone.

Comparison. Included experiments were required to have at least two comparison groups of filler similarity. The comparison groups in a single experiment could be, for example, lower and higher similarity; lower, mid, and higher similarity; unfair and fair similarity. Experiments were also required to have both target present and target absent conditions for joint consideration of the hit rate and false alarm rate and for discriminability to be calculated.

Outcomes. Included experiments were required to have reported counts or proportion (or percentage) of identification responses to perpetrators, innocent suspects, fillers, and lineup rejections. This is so that patterns in identification responses could be interpreted across experiments, and so ability to discriminate between innocent and guilty suspect (d') could be calculated for each similarity condition.

Quality assessment

Thirty eight experiments, that appeared to meet eligibility criteria after the first screening, were subject to a quality assessment using the critical appraisal skills program (CASP) randomised control trial appraisal tool (Public Health Resource Unit, 2006; see Appendix B). Experiments were assessed and scored for methodological quality, using the questions within the CASP tool. This considered factors such as the allocation to intervention groups, researchers being "blind" to the conditions and sample size. Each question could be answered as "yes", "can't tell" or "no". To create a scoring system, numerical value was added so that "yes" resulted in a score of 2, "can't tell" scored 1 and "no" scored 0. Questions 8 and 9 of the CASP tool were not used because they were open questions that could not be quantitively measured. Therefore, the highest total quality score that experiments could achieve was 16 and experiments that scored 7 or below were excluded on the basis of low methodological quality. Further details of individual article scoring, and quality assessment are provided in Appendix C.

Following the quality assessment process, seventeen experiments in total were excluded due to either low methodological quality (Devonport, & Cutler, 2004); use of secondary data sources (Booth, 2019; Cohen et al., 2020; Levi, 201); lack of reported data for discriminability analysis (Clark, & Tunnicliff, 2001; Cutler et al., 1987; Darling et al., 2008; Murray, & Wells, 1982); or experimental manipulations that did not fit within the scope of this review (Charman et al., 2011; Flowe et al., 2014; Gonzalez et al., 1995; Lindsay et al., 1991; Lindsay et al., 1994; Nosworthy, & Lindsay, 1990; Read et al., 1990; Tredoux et al., 2007; Wood, 2017). Quality scores of the remaining 21 experiments to be included in the review ranged from 10–16 (62.5%–100%; see Appendix C).

Data Extraction

The remaining 21 experiments were subject to data extraction using a data extraction form (see Appendix D). The data were categorised according to similarity comparisons in each experiment. Note however, that similarity manipulations are relative within each experiment, such that a high similarity condition in one experiment would not necessarily be comparable to a high similarity condition in another experiment. Therefore, relative definitions were used to categorise similarity manipulations in each experiment (e.g., lower vs higher similarity conditions; see Table 1). Descriptive information was collected about each experiment including how similarity was manipulated; lineup type (simultaneous or sequential presentation); medium (photo, video, other); target absent (TA) and target present (TP) conditions; if the fillers were suspect matched or perpetrator matched (i.e., if the same fillers were used across TA and TP conditions); how the innocent suspect was selected (e.g., description matched, or appearance matched) and any other information that was deemed to be relevant. Then the identification response data were recorded for each similarity condition type (proportion of suspect identifications, filler identifications, and lineup rejections, and also "don't know" responses if there was a "don't know" response option in a particular experiment).

Table 1

Similarity definitions used to code experimental conditions

Category	Definition
Unfair	*A lineup where someone who had not seen the perpetrator would be able to identify the suspect from the lineup on the basis that the suspect stands out as being different in physical appearance to the other lineup members.*
Fair	*A lineup where someone who had not seen the perpetrator would likely not be able to identify the suspect from the lineup at a rate higher than chance, because the*

	suspect does not stand out as being different in physical appearance to the other lineup members.
Lower	*Relative to other conditions within the experiment, fillers are less similar to the suspect (i.e., perpetrator or innocent suspect, depending on suspect matching procedure used), however the lineup appears fair (e.g., the suspect does not appear to stand out to someone who does not have a memory of the perpetrator.)*
Moderate	*Relative to other conditions within the experiment, fillers are of moderate similarity to the suspect, however the lineup appears fair.*
Higher	*Relative to other conditions within the experiment, the fillers are of higher similarity to the suspect, however the lineup remains fair.*

Dual coding

A second reviewer completed the data extraction process to allow for dual coding of the data and ensure research was interpreted objectively. The first and second reviewer completed the data extraction process individually using the similarity definitions and then met to compare data extracted. Discrepancies were discussed between the first and second reviewer and project supervisor. Each party outlined reasoning for their decision, consulted relevant articles for further information, and agreed on the final data label or category.

Feature Matching Model (Colloff et al., 2021) Hypotheses

Hypotheses about the hit rate and false alarm rate were made using theoretical predictions from the feature matching model (Colloff et al., 2021). Hypotheses were made following data extraction and prior to discriminability analysis. Details of hypotheses for each experiment are available in Table 2. Upon interpreting the data, z-tests were conducted for each of the similarity comparisons within experiments to examine if any differences across conditions were statistically significant (at a significance of $p < .05$). The z-tests conducted were one-tailed when there was a directional hypotheses (i.e., HR will increase as similarity decreases). When there was a null hypothesis (i.e., the FAR will remain unchanged), two-tailed z-tests were conducted. The outcome of z-tests was used to interpret

whether predictions of the feature matching model (Colloff et al., 2021) were supported within the literature. Results are displayed in Table 3.

Discriminability

Hypotheses about ability to discriminate between innocent and guilty suspects in each of the lineup conditions were made using theoretical predictions from diagnostic feature detection theory (Wixted & Mickes, 2014). Hypotheses were made following data extraction and prior to discriminability analysis. Details of hypotheses for each experiment are available in Table 2. Upon interpreting the data, the conceptual formula provided by Mickes et al. (2014); d' = z(correct ID rate) – z(false ID rate) was used to find out which lineup condition yielded better ability to discriminate between innocent and guilty suspects. This required the hit rate (HR) and false alarm rate (FAR) from each similarity condition to compute discriminability (d'), whereby a higher value indicates improved ability to discriminate between the innocent and guilty suspect. In experiments that reported proportions of identification responses for multiple experimental manipulations (e.g., suspect position), data were collapsed to calculate overall discriminability across conditions. d' values were interpreted considering theoretical predictions, and a d' value of 0 indicates an inability of witnesses to discriminate between innocent and guilty suspects (Macmillan & Creelman, 2004).

Results

See Table 2 for full details of reviewed experiments methodological characteristics and signal detection theory informed hypotheses of the impact of similarity on the HR, FAR and discriminability in each experiment. Suspect matched conditions are considered first, followed by perpetrator matched conditions. See Table 3 for lineup identification outcomes (i.e., HR and FAR), and a similarity condition comparisons. The results in Table 3 are reported in the order in which the hypotheses are presented in Table 2.

Table 2

Methodological characteristics and hypotheses of reviewed experiments (N= number of participants, L = Lower similarity lineups, M= Moderate similarity lineups, H= Higher similarity lineups, U= unfair similarity lineups, F= fair similarity lineup)

Author(s)	N	Similarity Conditions	Filler selection	Innocent suspect selection and similarity to perpetrator	Hypotheses
			SUSPECT MATCHED CONDITIONS		
Oriet & Fitzgerald (2018) ***EXPT 1***	415	L-M-H	Suspect matched. Selected by match to description and appearance. Single lineup paradigm.	Compared to fillers, highly similar to perpetrator, but not the most similar.	As similarity decreases, HR will increase, FAR will remain unchanged, *d′* will increase.
Oriet & Fitzgerald (2018) ***EXPT 2***	401	L-H	Suspect matched. Selected by match to appearance and description.	Compared to perpetrator and fillers, highly similar to perpetrator, but not the most similar.[1]	As similarity decreases, HR will increase, FAR will remain unchanged, *d′* will increase.
Colloff et al. (2016)	8925	U-F	Suspect matched. Selected by match to description.	Randomly selected from replication filler pool, matched to perpetrator description, and had a similar distinctive feature to perpetrator	In unfair compared to fair lineups, HR and FAR will increase, *d′* will decrease.

[1] Oriet & Fitzgerald (2018) similarity to perpetrator rating of 95th percentile, M= 3.90 (out of 10) of all filler to perpetrator ratings in both experiment 1 and experiment 2.

				that was digitally added using Photoshop CS5. In unfair lineups, most similar to perpetrator compared to fillers. In fair lineups, equally similar to perpetrator as fillers.	
Colloff et al. (2017)	2670	U-F	Suspect matched. Selected by match to description.	Randomly selected from replication filler pool, matched to perpetrator description, and had similar distinctive feature to perpetrator that was digitally added using Photoshop CS5. In unfair lineups, most similar to perpetrator compared to fillers. In fair lineups, equally similar to perpetrator as fillers.	In unfair compared to fair lineups, HR and FAR will increase, d' will decrease.
Smith et al. (2022) ***EXPT 1***	1365	U-F (high similarity)	Suspect matched. Selected by match to description. In fair conditions, fillers were high similarity to the perpetrator as they had a replicated distinctive feature of the perpetrator.	Randomly selected from replication filler pool, matched to perpetrator description, and had similar distinctive feature to perpetrator that was digitally added using Photoshop CS5. In unfair lineups, most similar to perpetrator compared to fillers.	In unfair compared to fair lineups, HR and FAR will increase, d' will decrease.

				In fair lineups, equally similar to perpetrator as fillers.	
Fitzgerald et al. (2015)	271	L-H	Suspect matched. Created using Fantamorph software, whereby five faces selected on match to appearance, then morphed with the suspect and other faces.	Created using Fantamorph software, morphed with 50% of perpetrator face. Most similar to perpetrator than fillers.	As similarity decreases, HR will increase, FAR will increase, *d′* will decrease.
Lucas & Brewer (2021)	3596	L-M-H	Suspect matched.[2] Created using Fantamorph software, whereby five low similarity faces selected on match to description and appearance, then in moderate[3] and high[4] similarity conditions, they	Match to description, selected on the basis they could be "confused" with the perpetrator. Compared to fillers, moderately similar to the perpetrator.	As similarity decreases, HR will increase, FAR will remain unchanged, *d′* will increase.

[2] After morphing, filler similarity was manipulative relative to the innocent and guilty suspect. That is, in the lower similarity condition, fillers matched the guilty suspect description, but were low similarity to both the guilty and innocent suspect.
[3] Moderate similarity faces were 33% morphed with the suspect face.
[4] High similarity faces were 50% morphed with the suspect face

			were morphed with the guilty suspect.		
Moreland (2015) ***EXPT 2***	991	L-H	Suspect matched. Selected by match to appearance.	Compared to fillers, moderate similarity to perpetrator. [5]	As similarity decreases, HR will increase, FAR will remain unchanged, *d′* will increase.
Moreland (2015) ***EXPT 3***	3011	L-H	Suspect matched. Selected by match to appearance.	Compared to fillers, moderate similarity to perpetrator.	As similarity decreases, HR will increase, FAR will remain unchanged, *d′* will increase.
Colloff et al. (2021) ***EXPT 1***	10559	L-M-H	Suspect matched. Selected by match to appearance and description.	Compared to fillers , median similarity to perpetrator.	As similarity decreases, HR will increase, FAR will remain unchanged, *d′* will increase.
Oriet & Fitzgerald (2018) ***EXPT 3***	363	L-H	Suspect matched. Selected by match to appearance and description.	Compared fillers, moderately similar to perpetrator. [6]	As similarity decreases, HR will increase, FAR will remain unchanged, *d′* will increase

[5] Moreland (2015) innocent suspect similarity to perpetrator M=2.02, filler similarity to perpetrator varied from M=1.19 – M=2.20 and one filler was more similar to the perpetrator than the innocent suspect.

[6] Oriet & Fitzgerald (2018) selected innocent suspect from suspect- filler pairs that were rated by judges as 40%-53% similar

Horry & Brewer (2016) ***EXPT 3***	25	L-M-H	Suspect matched. Created using FaceGen Modeler 3.5 (Singular Inversions, Inc).	Created using FaceGen Modeler 3.5 (Singular Inversions, Inc). Compared to fillers, low similarity to perpetrator.[7]	As similarity decreases, HR will increase, FAR will decrease, *d′* will increase.
Horry & Brewer (2016) ***EXPT 4***	23	L-M-H	Suspect matched. Created using FaceGen Modeler 3.5 (Singular Inversions, Inc).	Created using FaceGen Modeler 3.5 (Singular Inversions, Inc). Compared to fillers, low similarity to perpetrator.	As similarity decreases, HR will increase, FAR will decrease, *d′* will increase.
Horry & Brewer (2016) ***EXPT 5***	32	L-M-H	Suspect matched. Created using FaceGen Modeler 3.5 (Singular Inversions, Inc).	Created using FaceGen Modeler 3.5 (Singular Inversions, Inc). Compared to fillers, low similarity to perpetrator.	As similarity decreases, HR will increase, FAR will decrease, *d′* will increase
Juslin et al. (1996)	256	L-H	Suspect matched. Selected by match to appearance and description.	Matched to perpetrator description. Similarity not known. [8]	As similarity decreases, HR will increase, FAR will remain unchanged, *d′* will increase.

[7] Horry and Brewer (2016), innocent suspect was from low similarity filler pool, this was a seventh generation from the perpetrator using the FaceGen Modeler 3.5 (Singular Inversions) software.

[8] On the basis of match to description methodology reported by Juslin et al. (1996), it is assumed that the innocent suspect is not more similar to fillers than other lineup members.

Tunnicliff & Clark (2000) ***EXPT 1***	182	L-H	Suspect matched. Selected by match to appearance and description.	Matched to perpetrator description. Similarity not known.[9]	As similarity decreases, HR will increase, FAR will remain unchanged, *d′* will increase.
Tunnicliff & Clark (2000) ***EXPT 2***	148	L-H	Suspect matched. Selected by match to appearance and description.	Most similar to the perpetrator, compared to other fillers.	As similarity decreases, HR will increase, FAR will remain unchanged, *d′* will increase.
PERPETRATOR MATCHED CONDITIONS					
Flowe & Ebbesen (2007)	294	L-H	Perpetrator matched. Created using FACES software.	Compared to fillers, highly similar to perpetrator.	As similarity decreases, HR will increase, FAR will increase, *d′* will decrease.
Oriet & Fitzgerald (2018) ***EXPT 2***	401	L-H	Perpetrator matched. Selected by match to appearance and description.	Compared to fillers, highly similar to perpetrator, but not the most similar.[10]	As similarity decreases, HR will increase, FAR will increase, *d′* will decrease.

[9] On the basis of match to description methodology reported by Tunnicliff and Clark (2000), it is assumed that the innocent suspect is not more similar to fillers than other lineup members.

[10] Oriet & Fitzgerald (2018) similarity to perpetrator rating of 95th percentile, M= 3.90 (out of 10) of all filler to perpetrator ratings.

Oriet & Fitzgerald (2018) ***EXPT 1***	415	L-M-H	Perpetrator matched. Selected by match to description and appearance. Single lineup paradigm.	Compared to fillers, highly similar to perpetrator, but not the most similar.	As similarity decreases, HR will increase, FAR will increase, *d′* will decrease.
Lucas et al. (2021) ***EXPT 1***	623	L-M-H	Perpetrator matched. Selected by match to description.	Compared to fillers, highly similar to perpetrator.	As similarity decreases, HR will increase, FAR will increase, *d′* will decrease.
Lucas et al. (2021) ***EXPT 2***	3011	L-M-H	Perpetrator matched. Selected by match to description.	Compared to fillers, high similarity to perpetrator.	As similarity decreases, HR will increase, FAR will increase, *d′* will decrease.
Gronlund et al (2009)	2529	U-F-M	Perpetrator matched, but different fillers used in target present and target absent lineups. Selected by match to description.[11]	Compared to fillers, most similar to perpetrator.[12]	In unfair compared to fair lineups, HR will increase, FAR will increase, *d′* will decrease.

[11] Note that Gronlund et al. (2009) displayed two images of the same perpetrator in lineups, guilty strong was taken on the same day as the video and guilty weak was a photo taken a few weeks later when the perpetrator had grown facial hair and the length of their hair had changed.

[12] Gronlund et al. (2009) used two innocent suspects, innocent strong and innocent weak. Both were rated as most similar to the perpetrator description from a pool of 28 faces. However innocent strong was picked more often from a lineup than innocent weak.

Key et al. (2015)	2411	U-F	Perpetrator matched, but different fillers used in target present and target absent lineups. Selected by match to description[13].	Compared to fillers, most similar to perpetrator.	In unfair compared to fair lineups, HR and FAR will increase, d' will decrease.
Wetmore et al. (2015)	1584	U-F	Perpetrator matched[14], but different fillers used in target present and target absent lineups. Selected by match to description.	Compared to fillers, most similar to perpetrator.[15]	In unfair compared to fair lineups, HR and FAR will increase, d' will decrease.
Carlson et al. (2008) ***EXPT 2***	619	U-F-L	Perpetrator matched, but different fillers used in target present and target absent lineups. Selected by match to description.	Compared to fillers, highly similar to perpetrator.	As similarity decreases and the lineup becomes more unfair, HR and FAR will increase, d' will decrease.
Bergold & Heaton (2018)	871	L-M-H	Perpetrator matched. Selected by match to description.	Compared to fillers, highly similar to perpetrator.	As similarity decreases, HR will increase, FAR will increase, d' will decrease

[13] Key et al. (2015) used stimuli from Gronlund et al. (2009)
[14] Wetmore et al. (2015) used stimuli from Gronlund et al. (2009)
[15] Wetmore et al. (2015) used two innocent suspects from Gronlund et al. (2009), named innocent strong and innocent weak. Both were rated as most similar to the perpetrator description from a pool of 28 faces. However innocent strong was picked more often from a lineup than innocent weak.

Key et al. (2017)	818	U-F	Perpetrator matched. Selected by match to description.	Compared to fillers, highly similar to perpetrator.	In unfair compared to fair lineups, HR and FAR will increase, d' will decrease.
Lindsay & Wells (1980)	96	U-F	Perpetrator matched. Selected by match to description.	Matched to perpetrator description. Compared to perpetrator and fillers, more similar to perpetrator.	In unfair compared to fair lineups, HR and FAR will increase, d' will decrease.
Oriet & Fitzgerald (2018) ***EXPT 3***	363	L-H	Perpetrator matched. Selected by match to appearance and description.	Compared to fillers, moderately similar to perpetrator.[16]	As similarity decreases, HR will increase, FAR will increase, d' will decrease.
Colloff et al. (2021) ***EXPT 2***	9173	L-M-H	Perpetrator matched. Selected by match to appearance and description.	Compared to fillers, median similarity innocent suspect.	As similarity decreases, HR will increase, FAR will increase, d' will decrease.
Brewer & Wells (2006)	1200	L-H	Perpetrator matched. Selected by match to description and appearance.	Matched to perpetrator description. Similarity not known.	As similarity decreases, HR will increase, FAR will increase, d' will decrease.[17]

[16] Oriet & Fitzgerald (2018) selected innocent suspect from suspect- filler pairs that were rated by judges as 40%-53% similar

[17] Tentative hypotheses based on perpetrator matched element of the design, due to limited methodological details provided by authors regarding innocent suspect similarity.

Table 3

Lineup identification outcomes and similarity condition comparison in reviewed experiments

Author(s)	Lineup Type[18]	*N*[19]		HR[20]	FAR[21]	*d′*	Lineup	HR *z* test[22]		FMM support ?[23]	FAR *z* test[24]		FMM Support ?	DFD support ?[25]
		TP	*TA*					*z*	*p*[26]		*z*	*p*		
SUSPECT MATCHED CONDITIONS														
Oriet & Fitzgerald	L	71	72	0.58	0.07	**1.68**[28]	L-M	0.24	.406	No	0.00	1.00	**Yes**	**Yes**
(2018) *EXPT 1*[27]	M	69	55	0.56	0.07	1.63	L-H	3.88	<.001*	**Yes**	0.24	.813	**Yes**	**Yes**
	H	58	65	0.24	0.06	0.85	M-H	3.65	<.001*	**Yes**	0.22	.824	**Yes**	**Yes**
Oriet & Fitzgerald	L	66	62	0.77	0.05	**2.38**	L-H	2.11	.017*	**Yes**	0.00	1.00	**Yes**	**Yes**
(2018) *EXPT 2*	H	67	68	0.60	0.05	1.90								

[18] L= Lower similarity lineups, M= Moderate similarity lineups, H= Higher similarity lineups, U= unfair similarity lineups, F= fair similarity lineups.
[19] N= number of participants, TP= Target present lineups, TA = Target absent lineups.
[20] HR= Hit Rate (The rate at which the guilty perpetrator was identified in target present lineups).
[21] FAR= False Alarm Rate (The rate at which the innocent suspect was identified in target absent lineups).
[22] One tailed z-tests were used to find out if there was a significant difference in hit rates between similarity conditions.
[23] FMM= Feature Matching Model (Colloff et al., 2021). Hypotheses for each experiment are displayed in Table 2.
[24] z-tests were conducted to test if there was a significant difference in false alarm rates between similarity conditions, one-tailed tests were conducted when it was predicted there would be a directional change in the FAR (i.e., increase or decrease) and two-tailed tests were conducted when it was predicted there would be no change in the FAR across similarity conditions.
[25] DFD = Diagnostic feature detection (Wixted & Mickes, 2014) Hypotheses for each experiment are displayed in Table 2.
[26] * indicates there was a statistically significant difference between similarity conditions.
[27] Oriet and Fitzgerald (2018) perpetrator matched conditions are included in the perpetrator matched section below.
[28] *d′* = discriminability (ability to discriminate between innocent and guilty suspects). Bold font indicates that ability to discriminate between innocent and guilty suspect was improved in this condition, i.e., d′ was highest.

Colloff et al. (2016)	U	1110	1017	0.57	0.36	0.53	U-F	16.88	<.001*	**Yes**	21.03	<.001*	**Yes**	**Yes**
	F[29]	3397	3401	0.29	0.09	**0.79**							
Colloff et al. (2017)[30]-	U-y	113	89	0.81	0.45	1.00	U-y-	6.83	<.001*	**Yes**	7.82	<.001*	**Yes**	**Yes**
	F-y	340	348	0.44	0.10	**1.13**	F-y							
	U-m	113	89	0.69	0.47	0.57	U-m-	5.16	<.001*	**Yes**	8.19	<.001*	**Yes**	**Yes**
	F-m	340	348	0.41	0.10	**1.05**	F-m							
	U-o	113	89	0.60	0.39	0.53	U-o-	5.92	<.001*	**Yes**	6.68	<.001*	**Yes**	**Yes**
	F-o	340	348	0.29	0.10	**0.73**	F-o							
Smith et al. (2022)	U-s	184	191	0.72	0.40	0.84	U-s-	6.62	<.001*	**Yes**	7.12	<.001*	**Yes**	**Yes**
EXPT 1[31]	F-s	170	180	0.37	0.08	**1.07**	F-s							
	U-w	159	179	0.65	0.47	0.46	U-w-	6.29	<.001*	**Yes**	5.01	<.001*	**Yes**	**Yes**
	F-w	163	140	0.30	0.20	**0.76**	F-w							
	U-s[32]	184	191	0.85	0.62	0.73	U-s-	6.72	<.001*	**Yes**	8.84	<.001*	**Yes**	**Yes**
	F-s	170	180	0.52	0.17	**1.00**	F-s							
	U-w	159	179	0.82	0.68	0.45	U-w-	7.54	<.001*	**Yes**	9.08	<.001*	**Yes**	**Yes**
	F-w	162	140	0.41	0.17	**0.73**	F-w							

[29] The fair comparison of Colloff et al. (2016) is collapsed over three conceptually similar conditions of fair lineups for distinctive suspects. There was no significant differences between each of the three lineup conditions and so data were collapsed for comparison within the present review.

[30] Age of participant was considered by Colloff et al., 2017 and data were not collapsed in this review due to age related differences in discriminability observed by the authors. Young (y) = 18-30 years , middle (m) = 31-59 years , older (o) = 60+years.

[31] Witnesses viewed mock crime video three times in strong memory conditions (represented as U-strong in unfair lineups and F-strong in fair lineups) and once in weak memory condition (represented as U-weak in unfair lineups , F-weak in fair lineups).

[32] After rejecting the lineup, mock witnesses were asked to choose the person who most resembled their memory for the perpetrator in forced choice conditions.

Fitzgerald et al. (2015)	L	32	33	0.44	0.30	0.37	L-H	2.42	<.001*	**Yes**	0.44	.290	No	No
	H	31	34	0.16	0.24	-0.29								
Lucas & Brewer (2021)[33]	L-2p	807	812	0.70	0.35	**0.91**[34]	L-2p-M-2p	0.91	.182	No	0.84	.399	**Yes**	**Yes**
	M-2p	792	786	0.68	0.33	0.91	L-2p-H-2p	3.78	<.001*	**Yes**	2.13	.033*	No	**Yes**
	H-2p	789	778	0.61	0.30	0.80	M-2p-H-2p	2.91	.002*	**Yes**	1.23	.202	**Yes**	**Yes**
	L-3p	779	804	0.69	0.35	**0.88**	L-3p-M-3p	2.51	.006*	**Yes**	2.56	.010*	No	**Yes**
	M-3p	790	785	0.63	0.29	0.86	L-3p-H-3p	6.10	<.001*	**Yes**	4.39	<.001*	No	**Yes**
	H-3p	787	811	0.54	0.25	0.77	M-3p-H-3p	3.63	<.001*	**Yes**	1.80	.072	**Yes**	**Yes**
	L-6p	796	810	0.59	0.26	**0.87**	L-6p-M-6p	2.82	.002*	**Yes**	2.39	.017*	No	**Yes**
	M-6p	812	816	0.52	0.21	0.86	L-6p-H-6p	6.02	<.001*	**Yes**	3.90	<.001*	No	**Yes**

[33] Lucas and Brewer (2021) manipulated lineup size, 2p = 2 person lineup, 3p = 3 person lineup, 6p = 6 person lineup.

[34] Lower similarity yielded marginally higher discriminability of $d' = 0.9097$, compared to $d' = 0.9076$ in the mid similarity condition.

	H-6p	811	819	0.44	0.18	0.76	M-6p-H-6p	3.23	<.001*	**Yes**	1.52	.126	**Yes**	**Yes**
Moreland (2015)[35] *EXPT 2*	L	228	228	0.50	0.29	0.55	L-H	1.93	.027**	No	0.00	1	**Yes**	No
	H	228	228	0.59	0.29	**0.79**								
Moreland (2015) *EXPT 3*[36]	L-sim	207	205	0.57	0.23	**0.91**	L-sim-H-sim	0.82	.207	No	0.24	.809	**Yes**	**Yes**
	H-sim	206	205	0.53	0.22	0.83								
	L-seq	205	205	0.48	0.27	0.58	L-seq-H-seq	2.88	.001*	**Yes**	2.98	.003	No	No
	H-seq	205	205	0.34	0.15	**0.63**								
Colloff et al. (2021) *EXPT 1*	L	1794	1729	0.64	0.05	**2.00**	L-M	1.84	.030*	**Yes**	0.00	1.00	**Yes**	**Yes**
	M	1761	1817	0.61	0.05	1.92	L-H	7.12	<.001*	**Yes**	0.00	1.00	**Yes**	**Yes**
	H	1708	1750	0.52	0.05	1.70	M-H	5.35	<.001*	**Yes**	0.00	1.00	**Yes**	**Yes**
Oriet & Fitzgerald (2018) *EXPT 3*	L	51	65	0.78	0.26	**1.42**	L-H	2.35	.009*	**Yes**	0.66	.507	Yes	**Yes**
	H	62	62	0.57	0.21	0.98								
Horry & Brewer (2016) *EXPT 3*	L	25	25	0.79	0.23	1.56	L-M	0.65	.258	No	0.25	.402	No	No
	M	25	25	0.86	0.26	**1.75**	L-H	1.32	.094	No	0.17	.434	No	**Yes**
	H	25	25	0.62	0.25	0.98	M-H	1.93	.027*	**Yes**	0.08	.468	No	**Yes**
Horry & Brewer (2016) *EXPT 4*	L	23	23	0.45	0.16	**0.86**	L-M	0.07	.473	No	0.18	.428	No	**Yes**
	M	23	23	0.44	0.18	0.78	L-H	0.06	.290	No	0.09	.464	No	**Yes**
	H	23	23	0.37	0.17	0.62	M-H	0.48	.314	No	0.09	.456	No	**Yes**

[35] Note the data are collapsed over suspect position, which was manipulated in the orginal experiment.

[36] In addition to manipulating filler similarity, Moreland (2015) Experiment 3 compared simultaneous (L-sim, H-sim) and sequential (L-seq, H-seq) lineup presentation.

Horry & Brewer (2016) *EXPT 5*[37]	L	32	32	0.49	0.14	1.07	L-M	0.80	.468	No	0.12	.453	No	No
	M	32	32	0.48	0.13	**1.08**	L-H	0.81	.210	No	0.00	.500	No	**Yes**
	H	32	32	0.39	0.14	0.80	M-H	0.73	.234	No	0.12	.453	No	**Yes**
Juslin et al. (1996)	L	192	64	0.52	0.09	**1.40**	L-H	1.57	.058	No	0.00	1.00	**Yes**	**Yes**
	H	192	64	0.44	0.09	1.19								
Tunnicliff & Clark (2000) *EXPT 1*	L	128	128	0.53	0.13	1.23	L-H	0.00	1.00	No	2.95	.002*	No	No
	H	128	128	0.53	0.03	**1.94**								
Tunnicliff & Clark (2000) *EXPT 2*	L	48	48	0.31	0.19	0.40	L-H	0.21	.417	No	0.00	1.00	**Yes**	No
	H	48	48	0.33	0.19	0.45								
PERPETRATOR MATCHED CONDITIONS														
Flowe & Ebbsen (2007)	L	73.5[38]	73.5	0.33	0.34	-0.03	L-H	0.51	.306	No	3.00	.001*	**Yes**	**Yes**
	H	73.5	73.5	0.37	0.13	**0.79**								
Oriet & Fitzgerald (2018) *EXPT 2*	L	66	58	0.77	0.38	1.04	L-H	2.11	.018*	**Yes**	2.53	.006*	**Yes**	**Yes**
	H	67	58	0.60	0.17	**1.21**								
Oriet & Fitzgerald (2018) *EXPT 1*	L	71	72	0.58	0.07	**1.68**	L-M	0.24	.406	No	0.00	.500	No	No
	M	69	55	0.56	0.07	1.63	L-H	3.88	<.001*	**Yes**	0.23	.410	No	No
	H	58	65	0.24	0.06	0.85	M-H	3.65	<.001*	**Yes**	0.22	.412	No	No

[37] All lineups, consisted of fillers that were created using a "genetic" function of FaceGen Modeller (Singular Inversions Inc) software, this included two types of moderate similarity lineups called "Medium-High" and "Medium-Low" (Horry & Brewer, 2016). These conditions were collapsed over within the data set to enable comparison of low, moderate, and high similarity in the present review.

[38] The authors (Flowe & Ebbesen, 2007) reported the total sample size as 294, but did not provide a breakdown of sample size in each condition, therefore it has been estimated for the purposes of this review.

Lucas et al. (2021) ***EXPT 1***	L	312	312.5	0.55	0.20	0.97	L-M	1.25	.105	No	1.65	.309	No	**Yes**
	M	313	310	0.50	0.15	1.04	L-H	1.75	.040*	**Yes**	3.10	<.001*	**Yes**	**Yes**
	H	311	310.5	0.48	0.11	**1.19**	M-H	0.50	.308	No	1.48	.069	No	**Yes**
Lucas et al. (2021) ***EXPT 2***[39]	L-sr	199	202	0.58	0.29	0.76	L-sr-M-sr	2.01	.02*	**Yes**	2.34	.01*	**Yes**	**Yes**
	M-sr	202	199	0.48	0.19	0.83	L-sr-H-sr	1.99	.023*	**Yes**	3.92	<.001*	**Yes**	**Yes**
	H-sr	196	197	0.48	0.13	**1.08**	M-sr-H-sr	0.00	.500	No	1.63	.052	No	**Yes**
	L-lr	305	301	0.30	0.27	0.09	L-lr-M-lr	2.26	.012*	**Yes**	1.72	.043*	**Yes**	**Yes**
	M-lr	310	297	0.22	0.21	0.03	L-lr-H-lr	3.75	<.001*	**Yes**	4.68	<.001*	**Yes**	**Yes**
	H-lr	295	308	0.17	0.12	**0.22**	M-lr-H-lr	1.24	.105	No	2.99	.001*	**Yes**	**Yes**

[39] In addition to the similarity manipulation, the authors (Lucas et al., 2021) also manipulated retention interval (the time between viewing a stimulus video and completing a lineup identification task). In the short retention interval conditions, participants completed the lineup immediately after viewing the stimulus video. In the long retention, participants completed the lineup identification task between 16 and 21 days later. In the present review these conditions are represented as; L-sr (low similarity lineup, short retention), M-sr (moderate similarity lineup, short retention), H-sr (high similarity lineup, short retention), L-lr (low similarity lineup, long retention), M-lr (moderate similarity lineup, long retention), H-lr (high similarity lineup, long retention).

Gronlund et al (2009)[40]	U-sim-w	37	52	0.38	0.35	**0.08**	U-sim-w-F-sim-w	2.91	.002*	**Yes**	2.12	.012*	**Yes**	No
	F-sim-w	53.5	54.5	0.12	0.17	**0.08**								
	M-sim-w	53.5[41]	49	0.10	0.12	-0.11	U-sim-w-M-sim-w	3.19	<.001*	**Yes**	2.71	.003*	**Yes**	No
	U-seq-w	49.5	48.5	0.33	0.18	**0.48**	U-seq-w-F-seq-w	0.54	.293	No	0.71	.239	No	No
	F-seq-w	51	56.5	0.28	0.13	0.25								
	M-seq-w	54	50	0.36	0.13	-0.21	U-seq-w-M-seq-w	0.32	.372	No	0.69	.246	No	No
	U-sim-s	58.5	57.5	0.84	0.58	**0.79**	U-sim-s-F-sim-s	1.52	.064	No	0.00	.500	No	No
	F-sim-s	51	52	0.72	0.58	0.39								
	M-sim-s	50.5	55	0.68	0.58	0.29	U-sim-s-M-sim-s	1.97	.025*	**Yes**	0.00	.500	No	No
	U-seq-s	61	48.5	0.71	0.62	**0.25**	U-seq-F-seq	1.95	.025*	**Yes**	4.22	<.001*	**Yes**	No
	F-seq-s	50	49	0.53	0.45	0.20								

[40] In addition to similarity, the authors manipulated lineup presentation as being either simultaneous (e.g. all lineup members presented at once) or sequential (lineup members presented one at a time). Within this table, this is represented as: U-sim (simultaneous unfair similarity conditions) F-sim (simultaneous fair similarity conditions). M-Sim (simultaneous moderate similarity conditions). U-seq (sequential unfair similarity conditions) F-seq (sequential fair similarity conditions) and M-seq (sequential moderate similarity conditions). In some lineups, the image of the guilty suspect (named guilty weak by the authors) was taken weeks after the video when the perpetrator had grown facial hair and his hair style and length differed. The innocent suspect is these lineups was named "innocent weak" by the authors due to being picked from target absent lineups less often than the "innocent strong" innocent suspect in lineups below. This is represented as 'w' (e.g. U-sim-w, F-seq-w). In some lineups, the image of the guilty suspect (named guilty strong by the authors) was taken on the same day as the stimulus video and in the same location. The innocent suspect was also named innocent strong due to being picked more often from target absent lineups than the other innocent suspect known as "innocent weak". This is represented as 's' (e.g. U-sim-s, F-seq-s).

[41] Note that the number of participants is estimated due to data from lineups being collapsed over suspect location in the lineup manipulation.

	M-seq-s	52	55	0.36	0.33	0.08	U-seq-M-seq	.373	<.001*	**Yes**	2.95	.002*	**Yes**	No
Key et al. (2015)	U	280	452	0.67	0.29	**0.99**	U-F	5.63	<.001*	**Yes**	5.19	<.001*	**Yes**	No
	F	264	486	0.42	0.15	0.83								
Wetmore et al. (2015)[42]	U-w	142	195	0.78	0.27	1.39	U-w-F-w	1.73	.042*	**Yes**	4.11	<.001*	**Yes**	**Yes**
	F-w	146	208	0.69	0.11	**1.72**								
	U-s	142	75	0.78	0.66	0.46	U-s-F-s	1.73	0.42*	**Yes**	2.66	.004*	**Yes**	**Yes**
	F-s	146	70	0.69	0.44	**0.76**								
Carlson et al. (2008) *EXPT 2*[43]	U-sim	51	59	0.71	0.64	0.19	U-sim-F-sim	4.04	<.001*	**Yes**	5.03	<.001*	**Yes**	**Yes**
	F-sim	51	49	0.31	0.16	**0.50**								
	L-sim	47	66	0.43	0.30	0.35	U-sim-L-sim	2.80	.003*	**Yes**	3.81	<.001*	**Yes**	**Yes**
	U-seq	52	46	0.46	0.33	0.34	U-seq-F-seq	0.51	.306	No	1.45	.074	No	**Yes**
	F-seq	49	50	0.41	0.20	**0.61**								
	L-seq	51	48	0.24	0.38	0.40	U-seq-L-seq	2.34	.010*	**Yes**	0.60	.274	No	**Yes**

[42] Wetmore et al. (2015) used two innocent suspects from Gronlund et al. (2009), named innocent strong (in this review, represented as U-s unfair lineups and, F-s in fair lineups) and innocent weak (in this review, represented as U-w in unfair lineups and, F-w in fair lineups). Both were rated as most similar to the perpetrator description from a pool of 28 faces. However innocent strong was picked more often from a lineup than innocent weak.

[43] In addition to similarity, the authors (Carlson et al., 2008) manipulated lineup presentation as being either simultaneous (e.g. all lineup members presented at once) or sequential (lineup members presented one at a time). Within this table, this is represented as: U-sim (simultaneous unfair similarity conditions) F-sim (simultaneous fair similarity conditions). L-sim (simultaneous low similarity conditions). U-seq (sequential unfair similarity conditions) F-seq (sequential fair similarity conditions) and L-seq (sequential low similarity conditions).

Bergold & Heaton (2018)	L	114	149	0.47	0.05	**1.57**	L-M	1.45	.074	No	0.00	.500	No	No
	M	141	147	0.38	0.05	1.35	L-H	1.95	.026*	**Yes**	1.05	1.48	No	No
	H	144	146	0.35	0.08	1.02	M-H	.053	.299	No	1.04	1.49	No	No
Key et al. (2017)	U	121	128	0.65	0.40	0.64	U-F	1.86	.031*	**Yes**	5.66	<.001	**Yes**	**Yes**
	F	162	136	0.54	0.10	**1.38**								
Lindsay & Wells (1980)	U	11	11	0.71	0.70	0.03	U-F	0.64	.262	No	1.83	.034*	**Yes**	**Yes**
	F	11	11	0.58	0.31	**0.70**								
Oriet & Fitzgerald (2018) ***EXPT 3***	L	51	66	0.78	0.25	1.45	L-H	2.35	.009*	**Yes**	2.38	.008*	**Yes**	**Yes**
	H	62	50	0.57	0.08	**1.58**								
Colloff et al. (2021) ***EXPT 2***	L	1555	1489	0.62	0.10	1.59	L-M	1.70	.044*	**Yes**	5.26	<.001*	Yes	**Yes**
	M	1534	1567	0.59	0.05	**1.87**	L-H	7.19	<.001*	**Yes**	7.89	<.001*	Yes	**Yes**
	H	1464	1564	0.49	0.03	1.86	M-H	5.49	<.001*	**Yes**	2.86	.002*	Yes	No
Brewer & Wells (2006)[44]	L-t	300	300	0.34	0.32	0.06	L-t-	1.52	.067	No	0.26	.397	No	**Yes**
	H-t	301	299	0.40	0.33	**0.19**	H-t							
	L-w	300	300	0.66	0.52	**0.36**	L-w-	2.26	.019*	**Yes**	1.48	.070	No	**Yes**
	H-w	299	301	0.57	0.58	-0.03	H-w							

[44] In the stimulus video, participants viewed both a 'thief' and a 'waiter, and in subsequent lineups, the authors (Brewer & Wells, 2006) manipulated whether the suspect was the thief or the waiter and in target absent lineups, fillers were matched to the suspect (I.e., thief or waiter, depending on the condition). L-t= low similarity lineup, thief suspect, H-t = high similarity lineup, thief suspect. L-w= low similarity lineup, waiter suspect, H-w = high similarity lineup, waiter suspect.

Methodological Characteristics

The present review included twenty one papers published between 1980 and 2022. Those papers included twenty nine experiments that manipulated lineup filler similarity and tested 57,293 participants. Experiments utilised a photo or computer generated medium and lineup sizes ranged from two to eight. Seventeen experiments presented lineups simultaneously (Bergold & Heaton, 2018; Brewer & Wells, 2006; Colloff et al., 2016, 2017, 2021; Fitzgerald et al., 2015; Horry & Brewer, 2016; Juslin et al., 1996; Key et al., 2015, 2017; Lindsay & Wells, 1980; Lucas et al., 2021; Lucas & Brewer, 2021, Oriet & Fitzgerald, 2018; Smith et al., 2022, Tunnicliff & Clark, 2000; Wetmore et al., 2015). The remaining four experiments tested both simultaneous and sequential lineup presentation (Carlson et al., 2008; Flowe & Ebbesen, 2007; Grondlund et al., 2009; Moreland, 2015).

Similarity manipulation

Experiments were categorised according to similarity manipulations; lower, moderate, higher, unfair, or fair. Ten experiments compared lower, moderate, and higher similarity conditions (Bergold & Heaton, 2018; Colloff et al., 2021 *Experiment 1-2*; Horry & Brewer, 2016 *Experiment 3-5*; Lucas et al., 2021 *Experiment 1-2*; Lucas & Brewer, 2021; Oriet & Fitzgerald, 2018 *Experiment 1*). Eleven experiments compared lower and higher similarity conditions (Brewer & Wells, 2006; Carlson et al., 2008; Fitzgerald et al., 2015; Flowe & Ebbesen, 2007; Juslin et al. 1996; Moreland, 2015 *Experiment 2-3*; Oriet & Fitzgerald, 2018 *Experiment 2-3;* Tunnicliff & Clark, 2000 *Experiment 1-2*). There were nine experiments that compared fair and unfair lineups (Carlson et al., 2008 *Experiment 2*; Colloff et al., 2016, 2017; Gronlund et al., 2009; Key et al., 2015, 2017; Lindsay & Wells, 1980; Smith et al., 2022; Wetmore et al., 2015).

Filler Selection

Seventeen experiments used suspect matched filler selection methods (Colloff et al., 2016, 2017; Colloff et al., 2021 *Experiment 1*; Fitzgerald et al., 2015; Horry & Brewer, 2016 *Experiment 3*-5, Juslin et al., 1996; Lucas & Brewer, 2021; Moreland, 2015 *Experiment 2-3;* Oriet & Fitzgerald, 2018, *Experiment 1-3*; Smith et al., 2022, Tunnicliff & Clark, *Experiment 1-2*). Of the experiments that used suspect matched fillers, three experiments were matched by description (Colloff et al., 2016, 2017; Smith et al., 2022), three experiments were matched by appearance (Fitzgerald et al., 2015; Moreland, 2015 *Experiment 2-3*) and eight experiments used a combination of both match to description and appearance methods (Colloff et al., 2021, *Experiment 1*; Juslin et al., 1996; Lucas & Brewer, 2021; Oriet & Fitzgerald, 2018, *Experiment 1-3;* Tunnicliff & Clark, 2000 *Experiment 1-2*). A further three experiments used facial modelling software to create fillers (Horry & Brewer, 2016 *Experiment 3-5).*

Fifteen experiments used perpetrator matched filler selection methods (Bergold & Heaton, 2018; Brewer & Wells, 2006; Carlson et al., 2008 *Experiment 2*; Colloff et al., 2021 *Experiment 2*; Flowe & Ebbesen, 2007; Gronlund et al., 2009; Key et al., 2015, 2017; Lindsay &Wells, 1980; Lucas et al., 2021 *Experiment 1-2* ; Oriet & Fitzgerald *Experiment 1-3;* Wetmore et al., 2015). Of the experiments that used perpetrator matched fillers, nine experiments were matched by description (Bergold & Heaton, 2018; Carlson et al., 2008 *Experiment 2*; Grondlund et al., 2009; Key et al., 2015, 2017; Lindsay &Wells, 1980; Lucas et al., 2021 *Experiment 1*-2; Wetmore et al., 2015), five experiments used a combination of both match to description and appearance methods (Brewer & Wells, 2006; Colloff et al., 2021 *Experiment 2*; Oriet & Fitzgerald, 2018 *Experiment 1-3*). One experiment used facial modelling software (Flowe & Ebbesen, 2007). In none of the experiments were perpetrator matched fillers selected by appearance alone. Finally in four of the experiments that used perpetrator matched filler selection methods, different fillers were used in target present and

target absent conditions (Carlson et al., 2008 *Experiment 1* ; Gronlund et al., 2009; Key et al., 2015; Wetmore et al., 2015).

Innocent Suspect Selection

Experiments varied in how they selected an innocent suspect and how similar the innocent suspect was to the perpetrator compared to the other fillers. Seventeen experiments used an innocent suspect that was highly similar to the perpetrator (Bergold & Heaton, 2018; Carlson et al., 2008 *Experiment 2*; Colloff et al., 2016 *Unfair lineup;* Colloff et al., 2017 *Unfair lineup* Fitzgerald et al., 2015; Flowe & Ebbesen, 2007; Gronlund et al., 2009; Key et al., 2015, 2017; Lindsay & Wells, 1980; Lucas et al., 2021 *Experiment 1-2*; Oriet & Fitzgerald, 2018 *Experiment 1-3*; Smith et al., 2022 *Unfair lineup*; Wetmore et al., 2015). Ten experiments randomly selected a description matched innocent suspect (Brewer & Wells, 2006; Colloff et al., 2016 *Fair lineup*; Colloff et al., 2017 *Fair lineup*; Juslin et al., 1996; Oriet & Fitzgerald, 2018 *Experiment 1-3*; Smith et al., 2022 *Fair lineup*; Tunnicliff & Clark, 2000 *Experiment 1-2*). Innocent suspects of median or moderate similarity to the perpetrator were presented in five experiments (Colloff et al., 2021 Experiment *1-2*; Lucas & Brewer, 2021; Moreland, 2015 *Experiment 2-3*). Three experiments used an innocent suspect who was low similarity to the perpetrator (Horry & Brewer, 2016, *Experiment 3-5*).

Trends in the literature

As displayed in Table 2, predictions were made about the impact of methodological characteristics and similarity manipulations on the proportion of correct perpetrator identifications (HR), innocent suspect identifications (FAR) and ability to discriminate between innocent and guilty suspects (d'). These predictions were made using the feature matching model (Colloff et al., 2021) and diagnostic feature detection theory (Wixted & Mickes, 2014). The observed trends within the literature are displayed in Table 3 and will be described next.

Hit Rate (HR)

Theoretical predictions (see Table 2) about the pattern of the hit rate across conditions in a single in experiment were examined and z-tests were used to identify if any differences in hit rates were statistically significant (see Table 3). Out of 80 tests of the feature matching model (Colloff et al., 2021), predictions were supported by 65% (52). That is, the feature matching model (Colloff et al., 2021) was successful in predicting the effect of lineup conditions on identification patterns of the guilty suspect in 65% of similarity comparisons. However, results indicate that feature matching model (Colloff et al., 2021) did not account for findings in the remaining 35% (28) of similarity comparisons. To consider in what way predictions did and did not appear to be supported according to z-tests, lineup identifications at the experiment level are explored next.

Results indicate that eleven experiments supported feature matching model (Colloff et al., 2021) predictions (Colloff et al., 2016, 2017, 2021 *Experiment 1-2*; Fitzgerald et al., 2015; Key et al., 2015, 2017; Oriet & Fitzgerald, 2018 *Experiment 2-3*; Smith et al., 2022; Wetmore et al., 2015). That is, all target present lineup conditions followed the predicted pattern of results. Within the experiments that supported the feature matching model (Colloff et al., 2021), the hit rate increased significantly as similarity decreased in five experiments (Colloff et al., 2021 *Experiment 1-2*; Fitzgerald et al., 2015; Oriet & Fitzgerald, 2018 *Experiment 2-3*). And the hit rate increased significantly in unfair compared to fair lineups in six experiments (Colloff et al., 2016, 2017, Key et al., 2015, 2017; Smith et al., 2022; Wetmore et al., 2015).

Results indicate that eleven experiments partially supported feature matching model predictions (Colloff et al., 2021) as some target present lineup conditions followed the predicted pattern of results (Bergold & Heaton, 2018; Brewer & Wells, 2006; Carlson et al., 2008; Gronlund et al., 2009; Horry & Brewer, 2016 *Experiment 3*; Lucas & Brewer, 2021;

Lucas et al., 2021 *Experiment 1-2*; Moreland 2015 *Experiment 3;* Oriet & Fitzgerald, 2018 *Experiment 1-2).* In line with predictions, there was a significant difference in the hit rate for low and moderate similarity conditions in two experiments (Lucas et al., 2021 *Experiment 1-2*). In seven experiments, there was a significant difference in the hit rate for low and high similarity conditions (Bergold & Heaton, 2018; Brewer & Wells, 2006; Lucas & Brewer, 2021; Lucas et al., 2021 *Experiment 1-2*; Moreland 2015 *Experiment 3;* Oriet & Fitzgerald, 2018 *Experiment 1*). In three experiments, there was a significant difference in the hit rate for moderate and high similarity conditions (Horry & Brewer, 2016 *Experiment 3*; Lucas & Brewer, 2021; Oriet & Fitzgerald, 2018 *Experiment 1*). In three experiments, there was a significant difference in the hit rate for unfair and fair similarity conditions (Gronlund et al., 2009), unfair and medium similarity conditions (Gronlund et al., 2009) and unfair and low similarity conditions (Carlson et al., 2008).

In contrast to predictions, there was no significant difference in the hit rate for low and moderate similarity conditions in five experiments (Bergold & Heaton, 2018; Horry & Brewer, 2016 *Experiment 3*; Lucas & Brewer, 2021; Lucas et al., 2021 *Experiment 1*; Oriet & Fitzgerald, 2018, *Experiment 1*). In three experiments, there was no significant difference in the hit rate for low and high similarity conditions (Brewer & Wells, 2006; Horry & Brewer, 2016 *Experiment 3*; Moreland, 2015 *Experiment 3*). In three experiments, there was no significant difference in the hit rate for moderate and high similarity conditions (Bergold & Heaton, 2018; Lucas et al., 2021 *Experiment 1-2*). There were two experiments in which no significant difference in the hit rate of unfair and fair conditions were observed (Carlson et al., 2008 *Experiment 2*; Gronlund et al., 2009) and one experiment in which there were no significant difference in the hit rate of unfair and moderate lineup conditions (Gronlund et al., 2009).

Results indicate that seven experiments did not support the feature matching model (Colloff et al., 2021). That is, all target present lineup conditions did not follow the predicted pattern of results according to z-tests (Flowe & Ebbesen, 2007; Horry & Brewer, 2016 *Experiment 4-5*; Juslin et al., 1996; Lindsay & Wells, 1980; Moreland, 2015 *Experiment 2*; Tunnicliff & Clark, 2000). Within the experiments that did not support the feature matching model (Colloff et al., 2021), there was no significant difference in the hit rate across conditions of six experiments (Flowe & Ebbesen, 2007; Horry & Brewer, 2016 *Experiment 4-5*; Juslin et al., 1996; Lindsay & Wells, 1980; Tunnicliff & Clark, 2000 *Experiment 1-2*). That is, manipulating suspect-filler similarity did not appear to have a significant effect on the rate in which the guilty suspect was identified within six experiments. And in Moreland (2015 *Experiment 2*), there was a significant difference between low and high similarity conditions, but this was in the opposite direction than predicted, as the hit rate was highest in the high similarity condition.

False Alarm Rate (FAR)

Theoretical predictions (see Table 2) about the pattern of the false alarm rate across conditions within a single experiment were examined and z-tests were used to find out if any recorded differences within the data were statistically significant (See Table 3). Out of 80 tests of the feature matching model (Colloff et al., 2021), predictions were supported by 57.5% (46). That is, the feature matching model (Colloff et al., 2021) was successful in predicting the effect of lineup conditions on identification patterns of the innocent suspect in 57.5% of similarity comparisons. However, results indicate that feature matching model (Colloff et al., 2021) did not account for findings in the remaining 42.5% (34) of similarity comparisons. To consider in what way predictions did and did not appear to be supported according to z-tests, lineup identifications at the experiment level are explored next.

Results indicate that fifteen experiments supported feature matching model (Colloff et al., 2021) predictions (Colloff et al., 2016, 2017, 2021 *Experiment 1-2*; Flowe & Ebbesen, 2007; Juslin et al., 1996; Key et al., 2015, 2017; Lindsay & Wells, 1980; Oriet & Fitzgerald, 2018, *Experiment 1-3,* Smith et al., 2022; Tunnicliff & Clark, 2000 *Experiment 2*; Wetmore et al., 2015). That is, all target absent lineup conditions followed the predicted pattern of results. Within the experiments that supported the feature matching model (Colloff et al., 2021), the false alarm rate significantly increased as similarity decreased in three experiments (Colloff et al., 2021 *Experiment 2*; Flowe & Ebbesen, 2007; Oriet & Fitzgerald, 2018 *Experiment 2 perpetrator matched*). And the false alarm rate significantly increased in unfair compared to fair lineups in seven experiments (Colloff et al., 2016, 2017, Key et al., 2015, 2017; Lindsay & Wells, 1980; Smith et al., 2022; Wetmore et al., 2015). Finally, the false alarm rate remained unchanged in six experiments (Colloff et al., 2021 *Experiment 1*; Juslin et al., 1996; Oriet & Fitzgerald, 2018, *Experiment 1-3 suspect matched lineups,* Tunnicliff & Clark, 2000 *Experiment 2*). Note that "unchanged" refers to the confirmation of the null hypotheses, i.e., that there was no significant differences in the false alarm rates of similarity conditions.

Results indicate that six experiments partially supported feature matching model predictions (Colloff et al., 2021) as some target absent lineup conditions followed the predicted pattern of results (Carlson et al., 2008; Gronlund et al., 2009; Lucas & Brewer, 2021; Lucas et al., 2021 *Experiment 1*; Moreland, 2015 *Experiment 2-3*). In line with predictions, there was no significant difference in the false alarm rate for low and moderate similarity conditions in one experiment (Lucas & Brewer, 2021; Lucas et al., 2021 Experiment 2). In one experiment, there was no significant difference in the hit rate for moderate and high similarity conditions (Lucas & Brewer, 2021). There was no significant difference in the false alarm rate of low and high similarity conditions in four experiments

(Lucas et al., 2021 *Experiment 1-2;* Moreland 2015 *Experiment 2-3)*. In two experiments, there was a significant difference in the false alarm rate between unfair and fair similarity conditions (Carlson et al., 2008 *Experiment 2*; Gronlund et al., 2009). Finally, in one experiment there was a significant difference in the false alarm rate of unfair and low similarity conditions (Carlson et al., 2009).

In contrast to predictions, there was no significant difference in the false alarm rates of low and moderate similarity conditions in six experiments (Horry & Brewer, 2016 *Experiment 3-5*; Lucas & Brewer, 2021, Lucas et al., 2021 *Experiment 1*; Oriet & Fitzgerald 2018, Experiment *1 perpetrator matched*). In six experiments, there was no significant difference in the false alarm rate of low and high similarity conditions (Fitzgerald et al., 2015; Horry & Brewer, 2016 *Experiment 3-5;* Moreland, 2015; Oriet & Fitzgerald 2018, Experiment *1 perpetrator matched*; *1*). In six experiments, there was no significant difference in the false alarm rate of moderate and high similarity conditions (Horry & Brewer, 2016 *Experiment 3-5;* Lucas et al., 2021 *Experiment 1-2;* Oriet & Fitzgerald 2018, Experiment *1 perpetrator matched*). In two experiments, there was no significant differences in the false alarm rate of the unfair and fair lineups (Carlson et al., 2009 *Experiment 2*; Gronlund et al., 2009). Finally in one study there was no significant difference in the false alarm rate in unfair and medium similarity lineups (Gronlund et al., 2009) and in one study there was no significant difference in the false alarm rate of unfair and low similarity lineups (Carlson et al., 2008 *Experiment 2*). Finally, in two studies the difference in the false alarm rate was statistically significant when there was predicted to be no difference (Lucas & Brewer, 2021; Tunnicliff & Clark, 2000 *Experiment 1*).

Results indicate that eight experiments did not support the feature matching model according to z-tests (Colloff et al., 2021). That is, all target absent lineup conditions did not follow the predicted pattern of results (Bergold & Heaton, 2018; Brewer & Wells, 2007,

Fitzgerald et al., 2015; Horry & Brewer, 2016 *Experiment 3-5;* Oriet & Fitzgerald, 2018 *Experiment 1;* Tunnicliff & Clark, 2000 *Experiment 1*). Within the experiments that did not support the feature matching model (Colloff et al., 2021), there was no significant difference in the false alarm rate across conditions of seven experiments (Bergold & Heaton, 2018; Brewer & Wells, 2007; Fitzgerald et al., 2015; Horry & Brewer, 2016 *Experiment 3-5;* Oriet & Fitzgerald, 2018 *Experiment 1*). That is, manipulating suspect-filler similarity did not appear to have a significant effect on the rate in which the innocent suspect was identified within seven experiments. Finally, Tunnicliff and Clark (2000 *Experiment 1*) the false alarm rate was significantly higher in the low similarity condition compared to the high similarity condition. That is, in low similarity conditions the innocent suspect was picked from the lineup more often than in high similarity conditions.

Discriminability

Theoretical predictions (see Table 2) about the ability to discriminate between innocent and guilty suspects were examined using the conceptual formula provided by Mickes et al. (2014); *d'* = z(correct ID rate) – z(false ID rate) to compute discriminability (*d'*), whereby a higher value indicates a better ability for the witness to discriminate between the innocent and guilty suspect (see Table 3). Out of the 80 tests of diagnostic feature detection theory (Wixted & Mickes, 2014), predictions were supported by 71% (57). That is, diagnostic feature detection theory (Wixted & Mickes, 2014) was successful in predicting the effect of lineup conditions on ability to discriminate innocent from guilty suspects in 71% of similarity comparisons. However, results indicate that diagnostic feature detection theory (Wixted & Mickes, 2014) did not account for findings in the remaining 29% (23) of similarity comparisons. To consider in what was predictions were and were not supported, lineup identifications were explored at the experiment level next.

It was predicted that discriminability would increase as similarity decreased in thirteen experiments (Colloff et al., 2021 *Experiment 1*; Horry & Brewer, 2016 *Experiment 3-5*; Juslin et al., 1996; Lucas & Brewer, 2021; Moreland, 2015 *Experiments 2-3*; Oriet & Fitzgerald, 2018 *Experiment 1-3*; Tunnicliff & Clark, 2000 *Experiment 1-2*). This was supported across all conditions of eight experiments (Colloff et al., 2021 *Experiment 1*; Horry & Brewer, 2016 *Experiment 3-4*; Juslin et al., 1996; Lucas & Brewer, 2021; Oriet & Fitzgerald, 2018 *Experiment1, 2-3*). Results of Horry and Brewer (2016 *Experiment 3 & 5*) and Moreland (2015 *Experiment 3*) partially supported our hypotheses. In Horry and Brewer (2016 *Experiment 3-5*) discriminability increased as similarity decreased from higher to moderate similarity as predicted, but not from moderate to lower similarity conditions. Whereas in Moreland's experiment (*Experiment 3*), discriminability increased in the lower similarity condition when presented in a simultaneous lineup as predicted but decreased in the lower similarity condition when presented in a sequential lineup. Finally, three experiments did not support these predictions (Moreland, 2015 *Experiment 2*; Tunnicliff & Clark *Experiment 1-2*) as discriminability increased in higher similarity conditions.

It was predicted that discriminability would increase as similarity increased in fourteen experiments (Brewer & Wells, 2006; Bergold & Heaton, 2018; Colloff et al., 2021 *Experiment 2*; Fitzgerald et al., 2015; Flowe & Ebbesen, 2007; Horry & Brewer, 2016 *Experiment 3-5*; Lucas et al., 2021 *Experiment 1-2;* Oriet & Fitzgerald, 2018 *Experiments 1-3,* Tunnicliff & Clark, 2000 *Experiment 2*). Nine experiments supported the hypotheses as discriminability was increased in higher similarity conditions (Colloff et al., 2020 *Experiment 2;* Flowe & Ebbesen, 2007; Horry & Brewer, 201 *Experiment 4-5*, Lucas et al., 2021 *Experiment 1-2;* Oriet & Fitzgerald, 2018 *Experiment 2-3*; Tunnicliff & Clark, 2000 *Experiment 2*). Two experiments partially supported the hypotheses (Brewer & Wells, 2006; Horry & Brewer, 2016 *Experiment 3*). In Brewer and Wells (2006), the thief lineup led to

increased discriminability in the higher similarity condition as predicted, however the waiter lineup resulted in higher discriminability in the lower similarity condition. In Horry and Brewer (2016, *Experiment 3*), the moderate similarity condition led to the highest discriminability compared to lower and higher similarity conditions. Therefore this represents a partial replication, as it was predicted that discriminability would increase as similarity decreased, and this was the trend shown between high similarity and moderate similarity condition, but not between moderate similarity and low similarity conditions. Finally, three experiments displayed the opposite findings to hypotheses, whereby discriminability was higher in the lower similarity conditions (Bergold & Heaton, 2018; Fitzgerald et al., 2015; Oriet & Fitzgerald, 2018 *Experiment 1*).

Nine experiments included unfair conditions and it was predicted that discriminability would be lowest in these lineups (Carlson et al., 2008 *Experiment 2*; Colloff et al., 2016, 2017; Gronlund et al., 2009; Key et al., 2015, 2017; Lindsay & Wells, 1980; Smith et al., 2022; Wetmore et al., 2015). The pattern of results in seven experiments supported the hypotheses as discriminability decreased in unfair conditions compared to fair conditions (Carlson et al., 2008 *Experiment 2*; Colloff et al., 2016, 2017; Key et al., 2017; Lindsay & Wells, 1980; Smith et al., 2022; Wetmore et al., 2015). However, two experiments did not support predictions as discriminability was highest in unfair lineups (Gronlund et al., 2009; Key et al., 2015).

Discussion

Currently, there is an absence of scientific consensus on how best to select fillers to create fair lineups that protect the innocent and increase identifications of the guilty (Wells et al., 2020). The present review sought to explore optimal filler similarity through the lens of models based in signal detection theory: namely diagnostic feature detection theory (Wixted & Mickes, 2014) and the feature matching model (Colloff et al., 2021). The aims of the review were to examine how methodological characteristics of experiments influence experiment outcomes; how lineup filler similarity impacts identification of the perpetrator (hit rate), identification of the innocent suspect (false alarm rate); and witness ability to discriminate between innocent and guilty suspects. After applying inclusion/exclusion criteria, quality assessment and data extraction, twenty one papers presenting twenty nine experiments were included within the review (see Table 2 and Table 3).

Key Findings

There were no standardised procedures for constructing lineups within the similarity literature. This led to methodological variations in many aspects of the experiments, including filler selection (i.e., suspect or perpetrator matched, match to appearance or description or both), and innocent suspect selection (i.e., highly similar, description matched, moderately similar or dissimilar). Theoretically, these methodological variations influence the similarity between the fillers and the perpetrator; fillers and the innocent suspect; and the innocent suspect and the perpetrator. As a result, the methodological variations appear to influence patterns in the hit rate, false alarm rate and witness ability to discriminate between innocent and guilty suspects across similarity conditions. Consequently, it was not possible to distinguish a superior similarity level across all reviewed experiments. However, signal detection-based models (e.g., feature matching model; Colloff et al., 2021 and diagnostic-feature detection theory; Wixted & Mickes, 2014) offer a valuable insight into the impact of

methodological characteristics on the hit rate, false alarm rate and witness ability to distinguish between innocent and guilty lineup members. The models appear to predict patterns of results with reasonable success, and therefore can help to explain many of the contradictory findings in the literature to date.

Theoretical Underpinnings

It has been established that fair lineups, where the suspect does not stand out, are best practice (National Institute of Justice, 1999; Police and Criminal Evidence Act, 1984, Code D, 2011; Wells et al., 2020). However, unfair lineups should be avoided because they increase witness choosing, regardless of whether the suspect is innocent or guilty and make witnesses more likely to confuse innocent and guilty suspects (Clark, 2012; Colloff et al., 2016, 2017; Fitzgerald et al., 2013; Wells et al., 1979). Nevertheless, the methods to construct such fair lineups, including optimal similarity of fillers, is an area of research that is yet to be agree upon by academics and policymakers (Wells et al., 2020). Furthermore, the problem is further confounded as researchers do not routinely report the method in which they have constructed lineups (Quigley-McBride & Wells, 2021).

Previous reviews of the filler similarity literature have suggested that moderate and high similarity fillers are preferable to protect innocent suspects from misidentification (Fitzgerald et al., 2013, 2015). However, the impact of methodological characteristics (i.e., filler or innocent suspect selection) on witness identification ability has not previously been explored in reviews (see Fitzgerald et al., 2013, 2015). Furthermore, research in this field has used statistical tools which may have confounded the interpretation of experimental results (e.g., see Mickes et al., 2012). Therefore, a further review to identify the impact of lineup construction methodology on resulting suspect-filler similarity relations and witness lineup performance was conducted to establish what existing research can tell us about optimal filler similarity conditions.

To develop the fields' understanding of optimal filler similarity, it is helpful to apply a theoretical basis to the literature. The present review considered the filler similarity literature though the lens of signal detection theory (Wixted & Mickes, 2014) and complimentary theories of diagnostic feature detection hypotheses (Wixted, & Mickes, 2014) and feature matching model (Colloff et al., 2021). Signal detection theory assumes that when a witness views a lineup, the suspect (perpetrator or innocent suspect) and fillers generate a signal within a distribution in memory (Wixted & Mickes, 2014). The witness's ability to distinguish between innocent and guilty suspects is known as discriminability, and this is a measure of the degree of overlap between the memory distributions (Wixted & Mickes, 2014). Diagnostic feature detection theory describes the process in which witnesses make the identification decision. Specifically, witnesses discount features shared by all lineup members as they are nondiagnostic of guilt (Wixted & Mickes, 2014). However, features that are possessed only by the perpetrator and not by other lineup members will be diagnostic of guilt and can aid the witness in their identification accuracy (Wixted & Mickes, 2014).

The feature matching model develops this concept further by explaining the witness memory process at the time of the crime and how this influences identification response (Colloff et al., 2021). This model assumes that a face may be defined by its number of features and that each facial feature may have several possible settings (Colloff et al., 2021). For example, the feature of eye colour may have settings of brown, blue, hazel, grey and green. It follows that after witnessing a crime, the witness will have stored in memory, the unique features of the perpetrators face. When presented with a lineup in which the perpetrator is present, the encoded features of the perpetrator in the witnesses' memory will match those of the perpetrator presented in the lineup. However an innocent suspect and fillers in a lineup, who are not guilty, will not possess the same matching features as they are unique to the perpetrator, although there will be some overlap as the innocent suspect and

fillers will be matched to the perpetrator based on a small number of features identified in a witness' description (Colloff et al., 2021). Therefore, the features that the perpetrator possesses, and are not shared by the other lineup members, will be diagnostic of guilt and helpful for the witness in identification decision making (Colloff et al., 2021). Furthermore, the degree to which features are shared between the perpetrator, fillers and suspect will depend on the similarity of the lineup (Colloff et al., 2021). As such, filler selection methods resulting in dissimilar (i.e., low similarity) description matched fillers will enhance witness lineup identification performance, because the witness is able to use the perpetrators unique features that are not shared by other lineup members, and therefore diagnostic of guilt, to inform identification decision making (Colloff et al., 2021).

In agreement with others who highlight the importance of methodological decision-making in lineup construction (Quigley-McBride & Wells, 2021), signal detection theories make it clear that the methodological variances across the existing filler similarity literature are important in influencing the direction of results (i.e., which similarity condition is superior for witness identification). By applying the abovementioned theoretical assumptions (Wixted & Mickes, 2014; Colloff et al., 2021) and considering experimental methodology in previous studies, it was possible in the present review to predict outcomes in existing filler similarity literature for discriminability, correct perpetrator identifications (HR) and incorrect innocent suspect identifications (FAR) (see Table 2 for details).

HR and FAR

Predictions were made about the correct identification of the perpetrator (HR) and incorrect identification of the innocent suspect (FAR) using the feature matching model (Colloff et al., 2021). The predictions for target present conditions are simple. In target present conditions, when fillers have been description matched to the perpetrator, the HR should be higher in lower similarity than higher similarity conditions because there are fewer

shared features of lineup members and fewer fillers competing with the perpetrator in witness memory (Colloff et al., 2021). The predictions for target absent conditions are more complex, depending on the methodological choices in the experiment. In target absent conditions, when fillers and the innocent suspect were matched to the perpetrator description and filler similarity to the innocent suspect is manipulated, the FAR should not vary across similarity conditions. This is because changing the number of features that match across the fillers and the innocent suspect does not affect the number of features on the face of the innocent suspect that match witnesses' memory of the perpetrator. That is, theoretically the innocent suspect should not stand out in memory as being any more similar to the perpetrator than the other fillers in the lineup (Colloff et al., 2021).

However, in target absent conditions, where filler similarity is manipulated to the perpetrator and not the innocent suspect, it was predicted that the FAR would increase in lower similarity conditions. This is because the innocent suspect stands out in memory as being more similar to the perpetrator than the other fillers (Colloff et al., 2021). Similarly, when the innocent suspect is highly similar to the perpetrator compared to the fillers, it was predicted that the FAR would increase in low similarity conditions. This is because the innocent suspect stands out in memory as more similar to the perpetrator than the other fillers. Finally, when the innocent suspect is highly dissimilar to the perpetrator, the FAR was predicted to decrease in lower similarity conditions. This is because the innocent suspect does not stand out in memory as being any more similar to the perpetrator than the other fillers in the lineup.

Similarity conditions in the reviewed experiments were analysed to detect statistically significant differences, as a way to test feature matching model predictions (Colloff et al., 2021) for both the hit rate and false alarm rate. The feature matching model (Colloff et al., 2021) was successful in predicting the direction of the hit rate in 65% of target present

similarity comparisons. Furthermore, the feature matching model (Colloff et al., 2021) accurately predicted the direction of the false alarm rate in 57.5% of target absent similarity comparisons. Therefore, it is assumed that within these studies, the mechanisms of the feature matching model were accurate in predicating optimal filler similarity conditions (Colloff et al., 2021).

However, the feature matching model predictions were not supported by the z-tests on the hit rate in 35% of similarity comparisons and by the z-tests of the false alarm rate in 42.5% of similarity comparisons. This finding could suggest that the feature matching model could not account for witness identification in these lineups. Alternatively, it is possible that the lineup conditions that did not follow predictions do not counter earlier evidence of support for the feature matching model. Instead, unforeseen methodological factors in the original studies are likely to have impacted the pattern of results. These factors may not have been reported clearly by the authors in the original papers and were therefore not interpreted accurately during the data extraction process in the present review. The result of which means that the predicted pattern of results was not supported within these experiments. And it is likely that experimental outcomes were not in the direction predicted due to methodological factors, which will be described next.

One explanation for a conflicting pattern of results is the difficulty in operationalising similarity of experimental conditions, and this was also a difficulty in previous reviews of the filler similarity literature (Fitzgerald et al., 2013). In the present review, experimental lineups may have been labelled as low, moderate, or high similarity, however it is possible that this was not true within the corresponding experimental conditions. For example, in studies that observed no significant difference between low and moderate similarity conditions, and moderate and high similarity conditions, it is possible that the fillers were not dissimilar enough to warrant separate similarity categorisations. That is, the similarity categories

assigned by the initial researchers, or by the present review, may not have been an accurate representation of the actual similarity within the experimental lineups. This means that, feature matching model predictions may not have been accurate, because they did not consider the true filler similarity within lineup conditions. To overcome this problem, it is recommended that future research is transparent in reporting how similarity was manipulated and open access to experimental materials to enable replication of research and future reviews of the literature.

Furthermore, studies manipulated factors other than similarity, which may have confounded the results of the present review. Some studies manipulated lineup presentation, so that fillers were either presented one at a time (sequential presentation) or all at the same time (simultaneous presentation). The present review found that feature matching model predictions were not supported when lineups were presented sequentially (Carlson et al., 2008 *Experiment 2;* Gronlund et al., 2009; Moreland, 2015 *Experiment 3*). This suggests that presenting fillers one at a time (sequentially) may have impaired participants ability to identify unique features of the perpetrator that were diagnostic of guilt. And therefore presenting fillers all at the same time (simultaneously) was more helpful to enable participants to discount shared features of the lineup members that were not unique to the perpetrator and therefore not indicative of guilt.

However, the simultaneous lineup advantage was not consistent within the research as in Moreland (2015 *Experiment 3*), the hit rate predictions were supported by the sequential lineup and not the simultaneous lineup. Although a closer review of the data for Moreland (2015 *Experiment 3*) indicates that the hit rate was higher in all conditions of the simultaneous lineups compared to all conditions of the sequential lineups. This suggests that the simultaneous lineups did enable more identifications of the guilty suspect, but the effect of low similarity conditions did not appear to enhance suspect identification over high

similarity conditions. Therefore, it can be inferred that the impact of filler similarity on subsequent lineup identification outcome was influenced by lineup presentation (simultaneous or sequential), and this is also supported in previous literature reviews of filler similarity (Fitzgerald et al., 2013). More research on the impact of lineup presentation is needed to establish if there is clear interaction of simultaneous or sequential lineups and filler similarity on lineup identification outcomes.

Another factor that was manipulated within experiments was the memory strength of the suspect. That is, the strength of the memory signal for a suspect (i.e., strong, or weak). In Brewer and Wells (2006), the lineup that did not support predictions contained a different suspect (thief) to the lineup that did support predictions (waiter). From the description of the experimental conditions (Brewer & Wells, 2006), it is reported that within the stimulus video, the thief's face was viewed for 23 seconds whereas the waiters' face was viewed for 72 seconds. Therefore, the pattern of results may be explained by the reduction in exposure time that participants had to the thief's face. That is, there was no significant impact of manipulating similarity in the thief compared to waiter lineup, as the memory signal for the thief was weaker than for the waiter because participants had less time to encode the thief's facial features due to a shorter exposure time in the stimulus video. Therefore, length of exposure to the perpetrator appears to be another important methodological factor to consider in predicting optimal lineup conditions using the feature matching model (Colloff et al., 2021).

Other studies also manipulated memory strength of the suspect. In Grondlund et al. (2009) the strong memory lineups included an image of the guilty suspect which was taken on the same day as the stimulus video. Whereas in the weak memory lineups, the image of the guilty suspect was taken several weeks after the stimulus video, and the individual had grown facial hair and changed hairstyle. Moreover, in the innocent suspect conditions, the

'weak' innocent suspect was picked less often than the 'strong' innocent suspect. Furthermore, z-test results indicated that predictions were not supported when there was a weak guilty suspect and weak innocent suspect present in the lineup. However, in the predictions of the systematic literature review, the influence of manipulating suspect memory strength on filler similarity conditions and lineup identification was not considered.

When suspect memory strength is considered, the feature matching model would predict that in the target present strong memory lineups, there would be more unique features of the guilty suspect, that are not shared by other lineup members, that the participant is able to use to make an identification decision. However in the target present weak memory lineups, there would be less of the unique features encoded from the stimulus video as the guilty suspects' appearance has changed slightly and so the witness has less features from which to make an identification decision. Moreover, in the target absent lineups, the weak innocent suspect would have shared less features with the guilty suspect, and so was less likely to be identified from the lineup. Whereas the strong innocent suspect was likely to share more features with the guilty suspect, and so would be picked more often from the lineup. Therefore, it is argued that when the suspect is a poor match to the guilty suspect, then the effect of constructing an unfair lineup does not significantly increase the hit rate compared to a fair lineup. Therefore, the similarity in appearance between the perpetrator at the time of the crime and the perpetrator during the lineup appears to be another important methodological factor to consider in predicting optimal lineup conditions using the feature matching model (Colloff et al., 2021).

Another methodological factor manipulated by reviewed studies was lineup size. In Horry & Brewer (2016 *Experiment 3)* the lineup that did not support predictions consisted of the guilty suspect and one filler. As such, manipulating the similarity of a single filler did not significantly impact identification of the guilty suspect. Similarly, in Lucas and Brewer,

(2021) predictions were not fully supported in two and three person lineups. Theoretically, this makes sense because the witness has less faces in which to discount features that are non-diagnostic of guilt. Therefore, lineup size appears to be another factor that is highly relevant for predicting optimal lineup conditions using the feature matching model (Colloff et al., 2021).

However, lineup size may not be the only contributing factor to these findings. Within six person lineups, the predicted pattern of results was also not supported by Lucas and Brewer (2021). That is, there was a significant difference of the false alarm rate between similarity conditions (with a highest false alarm rate in the lower similarity condition), but there was predicted to be no change in the false alarm rate (Lucas & Brewer, 2021). When interpreting this study, the innocent suspect was coded as moderately similar to the perpetrator, however the observed pattern of results could suggest that the innocent suspect happened to be more similar to the perpetrator than the other fillers. Therefore, if the innocent suspect was highly similar to the perpetrator, then within a low similarity lineup, the innocent suspect would stand out from the other fillers and would therefore attract a higher rate of identifications. This further highlights how methodological characteristics, such as innocent suspect similarity, can influence lineup identification outcomes and therefore why it is difficult to apply feature matching model predictions based on the limited information reported in published papers, and without viewing the original stimulus materials.

Another confounding factor within the literature is the manipulation of similarity between fillers. For example, in Horry and Brewer (2016 *Experiment 4-5*), the similarity of lineup fillers was manipulated, so that there was one high similarity filler, and two fillers of 'high, medium-high, medium-low, or low' similarity to the suspect. This manipulation of similarity between the fillers themselves was not considered in application of feature matching model predictions as the overall similarity of fillers was coded into categories of

high, moderate, and low similarity to the suscept (or perpetrator) in the present review. However, upon revisiting this experiment, it is possible to predict that manipulating similarity within fillers would influence identification performance as some fillers will share more features with the guilty suspect than others, and so the participant is required to disentangle which features are unique to the perpetrator only and therefore indicative of guilt.

Moreover, the results of Horry and Brewer (2016) inform predictions about optimal lineup construction, as there is an effect of manipulating similarity within fillers, that appears to moderate the overall effect of suspect filler similarity on witness lineup identification. This highlights the need to establish a way to operationalise similarity categories (e.g. low, moderate, high) and the need to consider if all lineup fillers are from the same similarity category. Moreover, it suggests that the similarity between the fillers themselves, or the number of low, moderate, and high similarity filles is another important methodological factor to consider in predicting optimal lineup conditions using the feature matching model (Colloff et al., 2021).

Overall, the feature matching model (Colloff et al., 2021) was applied to the existing literature with reasonable success. That is, the hit rate and false alarm rate outcomes were accurately predicted in the majority of the literature. And in the studies that did not appear to support feature matching model predictions, it is argued that methodological characteristics (i.e., similarity category, lineup presentation, suspect memory strength, lineup size, similarity manipulation, innocent suspect similarity) are likely to have influenced the pattern of results. Next, findings are discussed in relation to diagnostic feature detection theory (Wixted & Mickes, 2014).

Discriminability

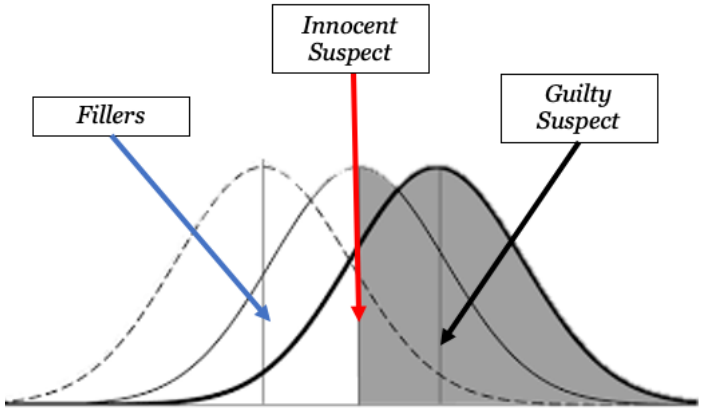
Diagnostic feature detection theory (Wixted & Mickes, 2014) predicts discriminability will change across lineup conditions according to methodological

characteristics of the experiment. This is, when making a lineup identification decision, optimal witnesses will discount shared lineup member features that are non-diagnostic (i.e., features that are not indicative of guilt because they are shared across all lineup members) and instead focus on diagnostic features that are not shared across all lineup members (Wixted & Mickes, 2014). Moreover, in high similarity lineups, the availability of potential diagnostic features is reduced due to the close resemblance between the fillers and perpetrator, and this results in reduced ability to discriminate between innocent and guilty suspects. Whereas in low similarity lineups, all lineup members will share some non-diagnostic features that have been used to match the fillers to the perpetrator or innocent suspect. However there are still diagnostic features available to make an identification decision, increasing witness ability to discriminate between innocent and guilty suspects.

Notably, the superiority of lower similarity conditions is dependent on lineups resulting in there being two memory distributions within witness memory, one for the perpetrator and the other for fillers and the innocent suspect (Colloff et al., 2021; Wixted & Mickes, 2014). Any methodological characteristic that impacts the memory distributions (i.e., creates three memory distributions, see figure 4) will subsequently impact ability to discriminate between innocent and guilty suspects and the superiority of the low similarity condition. Put another way, methodological characteristics that result in the innocent suspect standing out in memory, as being more similar to the perpetrator than the other fillers, will impair witness's ability to discriminate innocent from guilty suspects, particularly in low similarity lineups. Three methodological factors will be discussed next.

Figure 4

Three distribution model of lineup memory for fillers, innocent suspect, and guilty suspect in an unfair lineup, in which the innocent suspect stands out as more similar to the perpetrator than the fillers.



One important methodological factor is whether the fillers were suspect matched (i.e., fillers were matched to the perpetrator in target present conditions and to the innocent suspect in target absent conditions) or perpetrator matched (i.e., fillers were matched to the perpetrator in both target present and target absent conditions). If suspect matched fillers were presented, then theoretically there will be two memory distributions (i.e., the innocent suspect and the fillers are still in one memory strength distribution). In this case, the lower similarity advantage would remain because the innocent suspect does not stand out in memory as being any more similar to the perpetrator than the other fillers in the lineup (Colloff et al., 2021; Wixted & Mickes, 2014). That is, the witness is able to consider the low similarity fillers and innocent suspect, who share fewer unique features that match those of the perpetrator, meaning that the witness is more able to discount features that are non-diagnostic of guilt and as a result will be more able to discriminate between innocent and guilty suspects.

However, if target absent fillers are matched relative to the perpetrator the pattern of results is predicted to change. In a low similarity lineup, a third memory distribution is created for the innocent suspect, who now stands out from the other fillers as more similar to the witness's memory of the perpetrator (Colloff et al., 2021). In these conditions (where a third memory distribution exists), lower filler similarity lineups are no longer advantageous because the innocent suspect shares more features with the perpetrator compared to the fillers (Colloff et al., 20221; Wixted & Mickes, 2014). That is, the witness is less able to discount features that are non-diagnostic of guilt because the innocent suspect has more features that are shared with the perpetrator, and as a result the witness will be less able to discriminate between innocent and guilty suspects.

Another methodological characteristic predicted to impact discriminability was the selection of the innocent suspect. In conditions where the innocent suspect stood out from fillers, such as from being highly similar to the perpetrator when fillers were not, this would again create three memory distributions. As a result, there would be decreased ability to discriminate between innocent and guilty suspects within a lower similarity lineup (Colloff et al., 2021). Moreover, when the suspect (innocent or guilty) stands out, this would be considered to be an unfair lineup, and the trends in experiments that presented unfair lineups is presented next.

Theoretical predictions were applied to research on fair and unfair lineup conditions, where the suspect stands out as being more similar to the perpetrator than fillers. Consequently, three memory distributions are created, leading to decreased ability to discriminate between innocent and guilty suspects in unfair lineup conditions (e.g., Colloff et al., 2016). When reviewing the trends in existing literature, eight experiments appeared to support theoretical predictions made by diagnostic feature detection theory (Colloff et al., 2021, *Experiment 1*; Horry & Brewer, 2016 *Experiment 3- 4*; Juslin et al., 1996; Lucas &

Brewer, 2021; Oriet & Fitzgerald, 2018, *Experiment1-3)*. Furthermore, predictions were also supported in sixteen experiments that featured methodological characteristics leading to three memory distributions, as they resulted in improved ability to discriminate between innocent and guilty suspects in higher similarity conditions (Colloff et al., 2020, *Experiment 2;* Flowe & Ebbesen, 2007; Horry & Brewer, 2016, *Experiments 4-5*, Lucas et al., 2021, *Experiments 1- 2;* Oriet & Fitzgerald, 2018, *Experiments 2- 3*; Tunnicliff & Clark, 2000, *Experiment 2)* and decreased ability to discriminate between innocent and guilty suspects in unfair conditions (Carlson et al., 2008 *Experiment 2*; Colloff et al., 2016, 2017; Key et al., 2017; Lindsay &Wells, 1980; Smith et al., 2022; Wetmore et al., 2015). These findings support the diagnostic feature detection model and suggest that the theoretical mechanisms can be helpful in describing the process involved when a witness is tasked with identifying a guilty suspect in a lineup.

However, not all experiments in the review appeared to support the theoretical predictions of diagnostic feature detection theory. Six experiments displayed the opposite of predicted effects on ability to discriminate innocent from guilty suspects (Bergold & Heaton, 2018; Fitzgerald et al., 2015; Moreland, 2015, *Experiment 2*; Oriet & Fitzgerald, 2018, *Experiment 1;* Tunnicliff & Clark, 2000, *Experiment 1-2*). While this suggests there may be other mechanisms involved in witness identification, a closer examination of these experiments may offer insight into why predictions were not supported. These experiments are examined more closely next.

In two experiments, predictions may not have been supported due to confounding factors on similarity manipulations (Bergold & Heaton, 2018; Fitzgerald et al., 2015). In Bergold and Heaton (2018), for example, fillers were selected from databases of different sizes to create different levels of similarity (i.e., fillers from a smaller database were rated as less similar). A decrease in discriminability was predicted in low similarity conditions due to

the use of perpetrator matched fillers (which typically make the innocent suspect stand out in memory), however the results found that the lower similarity condition was superior. Therefore, it appears that there was an effect of database size, and that this effect was more influential on the identification outcomes than the use of perpetrator matched fillers. This was previously suggested by Quigley-McBride and Wells (2021), who also argued that using a large database is likely to result in fillers that are highly similar to the perpetrator. And this results in reduced hit rates due to it being harder to distinguish between the perpetrator and fillers (Quigley-McBride & Wells, 2021).

Furthermore, it was predicted that discriminability would increase in high similarity conditions in Fitzgerald et al. (2015). This is because of the authors use of facial morphing software, which was predicted to make the innocent suspect more similar to the perpetrator than fillers. Therefore, resulting in three memory distributions (i.e., one for the fillers, innocent suspect, and guilty suspect). However, results showed that the lower similarity superior advantage remained. On closer examination, d' in the higher similarity condition was a minus number. When interpreting d' , any value of below zero is assumed to mean that individuals were unable to discriminate (Macmillan & Creelman, 2004). Therefore, in Fitzgerald and colleagues' experiment, it is possible that participants were unable to complete the task in the higher similarity condition. So, the lower similarity condition was favourable. Alternatively, it is possible that the innocent suspect was not more similar to the perpetrator than the other fillers, who were also morphed with the suspect face.

In two experiments, discriminability was predicted to increase in low similarity conditions. However, the opposite effect was observed, therefore suggesting there were three memory distributions (Moreland, 2015 *Experiment 2*, Tunnicliff & Clark, 2000 *Experiment 1*). In both experiments, the innocent suspect was described as being description matched, however it is possible that the innocent suspect was more similar to the perpetrator than other

fillers by chance, leading to three memory distributions that would make higher similarity conditions favourable (Colloff et al., 2021; Wixted & Mickes, 2014). Finally, in Oriet and Fitzgerald (2018 *Experiment 1*), the innocent suspect was predicted to be more similar to the perpetrator, leading to three memory distributions and a higher similarity condition advantage. However lower similarity conditions were superior. This suggests that the innocent suspect did not stand out as predicted, resulting in superiority in the lower similarity condition.

Limitations

Overall, it is difficult to know for certain whether some experiments did not support the hypotheses because of extraneous factors in the experiment manipulations and design (e.g., database size; computer generated stimuli; innocent suspect similarity to the perpetrator); misinterpretation of the experiment methods by the reviewer because the experiment materials could not be viewed, or the details were not reported; or due to an issue with diagnostic feature detection theory or the feature matching model. In the future, more work could test diagnostic feature detection theory and the feature matching model (Colloff et al., 2021; Wixted & Mickes, 2014) by developing a priori hypotheses and testing these in high-powered, well-designed primary experiments.

This review utilised a comprehensive data extraction process and dual coding process whereby two independent reviewers agreed on the extracted data and categorisation of similarity and experiment characteristics. This allowed for comparison of reviewer's interpretation of the literature to ensure that similarity definitions and categories were operationalised. As has been previously identified (Quigley-McBride & Wells, 2021), within existing research there is an absence of reporting of specific details of lineup methodological characteristics, such as filler and innocent suspect selection and presentation. In the present review, this made it difficult to make theoretical predictions about patterns of results. As

such, the findings of this review support Quigley-McBride and Wells (2021) recommendations that researchers thoroughly report lineup construction methods and make experimental materials and data publicly available.

Moreover, the limitations of the vote counting method utilised in the present review should be noted. That is, the present review methodology involved 'taking a vote' by reviewing how many results were significantly positive, significantly negative or if there was no significant relationship between variables (Bushman & Wang, 2009). This methodology is discouraged because it does not take into account individual study sample size, which can be problematic because a larger sample size can make it more likely that an effect will be found (Bushman & Wang, 2009). Moreover, the vote counting methodology also does not consider effect size, which is how meaningful the relationship of manipulated variables is (Bushman & Wang, 2009). As such, it is recommended that when there have been further studies that report methodological details and have sample sizes enabling experimental effects to be reliably detected, a meta-analysis should be conducted to further test optimal lineup conditions as a function of methodological decision making.

Conclusion

This review has examined trends within the existing filler similarity literature, combined with theoretical understanding from signal-detection based models (Colloff et al., 2021; Wixted & Mickes, 2014). It has been highlighted that methodological inconsistencies in research have contributed to the lack of academic consensus regarding optimal filler similarity conditions, and future research is required to disentangle the influence of methodological inconsistencies on lineup identification outcomes. Despite methodological limitations of the existing literature, this review has demonstrated that signal detection models can predict lineup identification outcomes with reasonable success, suggesting that the mechanisms of the feature matching model (Colloff et al., 2021) and diagnostic feature

detection theory (Wixted & Mickes, 2014) may, in the future, help to inform practice on constructing optimal lineup conditions. Although it has previously been argued that low similarity lineups increase the risk of innocent suspect misidentification (Fitzgerald et al., 2013), the present review highlights that low filler similarity conditions can be optimal for increasing witness's ability to discriminate innocent from guilty suspects. Moreover, the low similarity lineup advantage is possible without increasing the risk of misidentification of innocent suspects. As explained by diagnostic feature detection theory and the feature matching model, low similarity lineups allow the witness to focus on the perpetrators unique features that are diagnostic of guilt to make an accurate identification decision (Colloff et al., 2021: Wixted & Mickes, 2014). Importantly though, these optimal conditions can be achieved only when the innocent suspect does not stand out. That is, low similarity lineups are only beneficial to performance when using lineup construction methods that result in the innocent suspect being no more similar to the perpetrator than the other lineup members, such as avoiding the use of high similarity innocent suspects and using suspect matched filler selection methods in the target-absent conditions.

CHAPTER 3 : CONSTRUCTING LINEUPS FOR DISTINCTIVE SUSPECTS

Abstract

Constructing lineups for distinctive suspects can involve replicating the distinctive feature across the fillers, but research has not yet explored how similar the replicated feature should be to optimise witness performance. In an experiment (N=4915), this chapter compared unfair do-nothing lineups wherein the suspect stood out, to high similarity replication lineups where the fillers had a very similar distinctive feature as the suspect, and low similarity replication lineups where the fillers had a similar, but non-identical, feature to the suspect. Participants viewed a mock crime video before they were randomly allocated to one of six line up conditions (target present or target absent; high similarity replication, low similarity replication or do-nothing). Compared to unfair and high similarity replication lineups, low similarity replication lineups enhanced witness ability to discriminate between innocent and guilty suspects. Compared to high similarity lineups, low similarity lineups increased the hit rate of guilty suspects, without changing the false alarm rate to innocent suspects. These results are predicted by signal detection-based models of diagnostic feature detection (Wixted & Mickes, 2014) and feature matching model (Colloff et al., 2021) and suggest that replicating a similar but non-identical distinctive feature may optimise witness performance on lineups for distinctive suspects. Finally, this chapter develops theory further by highlighting the different processes in high similarity replication, low similarity replication, and do-nothing lineups.

Introduction

Eyewitness evidence is routinely used by the Criminal Justice System as a means of convicting the perpetrator of a crime. Within this practice, a witness of a crime is often asked to complete a police identification procedure known as a lineup. In a lineup, the witness is presented with the suspect (who police believe might be the perpetrator) and similar looking people known as fillers (who are individuals known to be innocent). An eyewitness is then asked to decide by stating if the perpetrator is present or absent within the lineup. If the witness makes an identification of the suspected perpetrator, this identification decision may then be used as evidence within a trial to convict the suspected perpetrator of a crime (Brewer & Wells, 2006).

However, the use of lineup procedures and eyewitness evidence in the Criminal Justice System can be vulnerable to witness error. One type of error occurs when an innocent suspect is mistakenly identified and at risk of conviction for a crime they did not commit. The real world implications of eyewitness error are apparent as eyewitness misidentifications were a factor involved in 69% of the 375 overturned convictions of innocent individuals by DNA evidence in the US (Innocence Project, 2022). Another type of error occurs when an eyewitness fails to correctly identify the perpetrator (i.e., the guilty suspect) from the lineup. Research indicates that witnesses are only 50% likely to identify a person they have seen before (Wells et al., 2006). Furthermore, a review of 94 lineup experiments found that 21.2% of mock witnesses identified a filler when the perpetrator was present, and 34.6% of mock witnesses made an identification when the perpetrator was not present (Clark et al., 2008). This highlights the susceptibility of witness identifications to inaccuracy and the need for further understanding of the processes involved in eyewitness identification to increase the correct identification of guilty suspects and protect the innocent from being misidentified.

Typically, researchers have investigated eyewitness identification using an experimental mock crime paradigm. This involves exposure of participants to a staged crime, either in real time or by video. Then participants, who are now “witnesses”, are asked to identify if the perpetrator is present in a lineup. Experimenters can manipulate the lineup conditions so that they are either “target present”, whereby the perpetrator of the mock crime is present within the lineup, or “target absent”, where the perpetrator of the mock crime is not present within the lineup and may be replaced by a designated innocent suspect or a randomly allocated filler. The use of target present and target absent conditions is an attempt to represent real life, where the true guilt or innocence of the suspect is not known and so lineups could be target present or target absent.

The outcomes of laboratory experiments have been categorised into factors influencing eyewitness responses, known as estimator and system variables (Wells, 1978). Estimator variables are the factors during the crime that may impact witness memory and system variables are the processes used by the Criminal Justice System, such as the way police investigate the crime. Experiments report estimator variables that can impact witness memory accuracy include encoding conditions (Lindsay et al., 1998), race of the perpetrator and witness (Chance & Goldstein, 1996), stress (Deffenbacher et al., 2004), and the presence of a weapon (Loftus et al., 1987). Experiments report system variables that can impact witness memory accuracy include retention interval (i.e. the time between witnessing a crime and completing a lineup identification procedure; Juslin et al., 1996; Palmer et al., 2013; Read et al., 1998), lineup presentation (i.e. whether lineup members are presented alongside each other, simultaneously, or one at a time, sequentially; Lindsay &Wells, 1985; NRC, 2014; Steblay et al., 2003; 2011); if the witness receives feedback about their memory (Eisen et al., 2008; Luus & Wells, 1994; Smalarz & Wells, 2014; Starzynski et al., 2005; Wade et al., 2018; Wells & Bradfield, 1998); and the impact of the instructions given to the witness

when completing a lineup identification task (Clark, 2005; Clark et al., 2014; Lindsay et al., 1991; Mickes et al., 2017; Steblay et al., 1997). Research has also considered the impact of how similar looking the lineup members are within the lineup (Fitzgerald et al., 2013, 2015; see chapter 2). Together, this research has been used to develop theoretical understanding of eyewitness memory and recommendations for practice to increase identification of guilty suspects and protect innocent suspects from misidentification (NRC, 2014; Technical Working Group For Eyewitness Evidence, 1999; Wells, 1998; Wells et al., 2020).

Following a review of the eyewitness literature, a key recommendation was to use lineup methods that enhance discriminability (NRC, 2014). This means constructing lineup procedures that result in increased identification of the guilty suspect (known as the Hit Rate, HR) and decreased misidentification of the innocent suspect (known as the False Alarm Rate, FAR). Procedures that enhance discriminability simultaneously increase the HR and decrease the FAR, compared to procedures that yield lower discriminability. When considering discriminability, it is important to distinguish between empirical and theoretical discriminability. Firstly, empirical discriminability is calculated statistically by the area under the ROC curve (AUC) and refers to the degree to which a witness is able to accurately sort innocent and guilty suspects into their respective groups (Wixted & Mickes, 2018). Whereas theoretical discriminability can be measured using the d-prime statistic (*d'*) and refers to the amount of theoretical overlap between the memory strengths for innocent and guilty suspects in the witness's memory (Wixted & Mickes, 2018). Experiments have considered both empirical and theoretical discriminability to identify which system and estimator variables contribute to a witness being able to discriminate between innocent and guilty suspects (Wixted & Mickes, 2018), and the results of those two types of analyses typically agree (but see Wilson et al. 2018; and Rotello & Chen, 2016 for exceptions).

Within the literature at present, a key area of debate is how to enhance both empirical and theoretical discriminability when constructing lineups for suspects with distinctive facial features, such as tattoos, scars, and bruising such as a black eye. It is estimated that up to a third of lineup suspects have distinctive facial features (Flowe et al., 2018). Constructing lineups for distinctive suspects can be considered a system variable, as police are required to consider how to accommodate distinctive suspects within lineups. Current policing practice requires an identification procedure to be fair, meaning that the suspect should not stand out because they look different to the other fillers (Police and Criminal Evidence Act, 1984, Code D, 2011). However, there are multiple methods to ensuring that a suspect with a distinctive facial feature does not stand out. The variability in methods used by law enforcement was highlighted by a survey of US police practice, which found that 77% of respondents attempted to replicate the distinctive feature onto fillers, of which 23% added a similar distinctive feature, and 18% attempted to conceal the feature across all lineup members (Wogalter et al., 2004). Moreover, 30% of participants reported that they did not do anything when suspects had distinctive facial features (Wogalter et al., 2004). The variability in construction of lineups for distinctive suspects may therefore contribute to variabilities in witness responses and ability to accurately identify the guilty suspect.

Similarly to the U.S, there is variability in the lineup construction methods for distinctive suspects in the United Kingdom (Police and Criminal Evidence Act 1984, Code D, 2011; Technical Working Group for Eyewitness Evidence, 1999). Lineups may be constructed using concealment techniques, such as placing a block or pixelation over the distinctive feature of the suspect and fillers in the same area (see figure 5). Another method involves replicating the distinctive feature across the other lineup members, and this is usually administered digitally. However, the implementation of the replication method is not standardised. For example, a replication lineup for a distinctive suspect with a tattoo on the

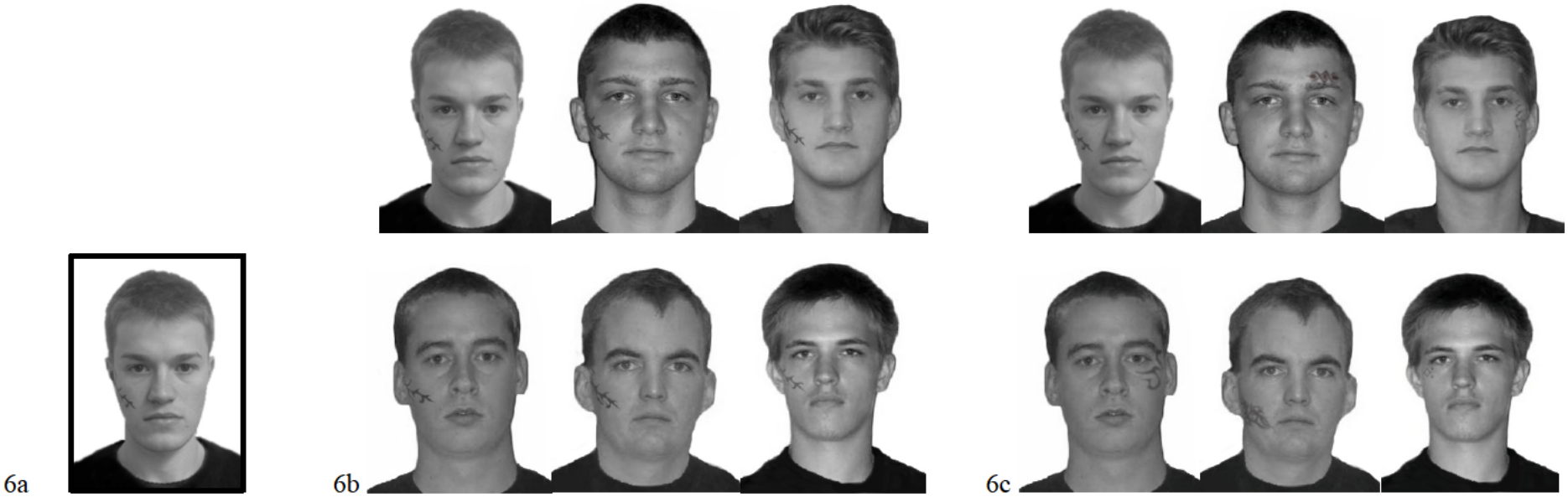
right cheek could include fillers with the exact same tattoo on the right cheek (high similarity replication), or the tattoo style and location could be varied across fillers (low similarity replication; see figure 6). How do the different methods of accommodating distinctive suspects influence eyewitness performance? I will discuss the existing research and theory, next.

Figure 5

Example of distinctive suspect (5a) and target present six person simultaneous lineups using pixelation (5b) and block (5c) concealment methods (from Colloff et al., 2016).

Figure 6

Example of distinctive suspect (6a) and target present six person simultaneous lineups using high similarity replication (6b) and low similarity replication (6c) lineup construction methods.



Early experiments of lineup construction techniques for suspects with distinctive facial features compared replication and removal methods (Badham et al., 2013; Zarkadi et al., 2009). In these experiments, replication lineup conditions involved the addition of the suspects distinctive feature to fillers. However, the removal method used by these experiments differed from police practice of pixelation or block methods. Instead, this condition involved participants viewing a suspect with a distinctive facial feature in the encoding stage of the experiment, and then at the identification stage, the suspect's distinctive feature had been removed and participants viewed a lineup in which neither the original distinctive suspect nor fillers had a distinctive facial feature. Both experiments found that participants made more perpetrator identifications in the replication compared to removal conditions, without increasing identifications of the innocent suspect in target absent conditions (Badham et al., 2013; Zarkadi et al., 2009). This research suggests that replication lineup conditions aid witnesses in correctly identifying the guilty suspect compared to when the feature is removed. However it was not possible to consider the impact on discriminability as neither experiment calculated empirical or theoretical discriminability across the lineup conditions.

More recently, Jones et al. (2020) investigated distinctive features by presenting participants with replication and removal lineups for distinctive suspects with a black eye or non-distinctive suspects without a black eye. Jones et al. (2020) reported that in distinctive suspect conditions (in which participants viewed a distinctive suspect at encoding and lineup conditions) there was no significant difference in discriminability between replication and removal lineup conditions. This suggests that directly replicating the distinctive feature was no more helpful than removing the distinctive feature to aid witness memory performance. However, the removal technique used by Jones et al. (2020) and previous research (Badham et al., 2013; Zarkadi et al., 2009) is not representative of current U.K police practice, and so

the comparison of replication and removal concealment methods may not be the most informative when the goal is to make policy and practice recommendations.

Further research has compared fair and unfair lineup construction methods for suspects with distinctive facial features (Colloff et al., 2016, 2017). In both experiments, lineups were considered fair when the distinctive suspect did not stand out from other fillers. This was achieved using replication (in which all lineup members had the same distinctive facial feature as the distinctive suspect), block and pixelation methods that reflected current U.K policing methods. In contrast, lineups were unfair when the distinctive suspect stood out, meaning that only the suspect (either innocent or guilty) had a distinctive feature and all other lineup members did not. Colloff et al. (2016) reported that unfair lineups led to an impairment in witness ability to accurately allocate guilty suspects and innocent suspects into their correct categories compared to fair lineups. This finding was replicated in Colloff et al. (2017), who tested witnesses of different ages, and again witnesses were less able to discriminate between innocent and guilty suspects in the unfair lineup conditions compared to fair conditions. In both experiments, there was no difference between the three fair lineup conditions: pixelation, block and replication (Colloff et al., 2016, 2017).

An experiment by Smith et al. (2022) compared a fair lineup that was equivalent to a high similarity replication condition (in which all lineup members had the same distinctive facial feature as the distinctive suspect) and an unfair lineup in which only the suspect (innocent or guilty) had the distinctive facial feature. They reported that fair lineups did not improve discriminability and argue that the advantage of fair lineups exists because they lead to a distribution of choices away from the suspect and onto fillers and this leads to improved discriminability of the outside observer, who is aware of which lineup members are fillers (Smith et al., 2022). However, the conclusions of this paper are likely due to the analysis technique used (see Wilson & Colloff, 2020 and Starns et al., 2022 for critiques of the

investigator discriminability approach). When the goal is to measure participants ability to discriminate between innocent and guilty suspects, traditional ROC analysis from the basic memory and perceptual literature or calculation of signal-detection based statistics (e.g., *d′* by fitting a theoretical model to the data) are recommended (NRC, 2014). When fitting a signal-detection model to all of the empirical data (suspects, fillers, rejects), Colloff et al. (2016, 2017, 2020), reported better ability to discriminate between innocent and guilty suspects in fair compared to unfair lineups. Regardless of the differences in analytical techniques used across experiments and conclusions about discriminability, researchers and practitioners agree that using fair lineup techniques such as replication, pixelation or block concealment methods is desirable over using do-nothing (unfair) lineups for distinctive suspects.

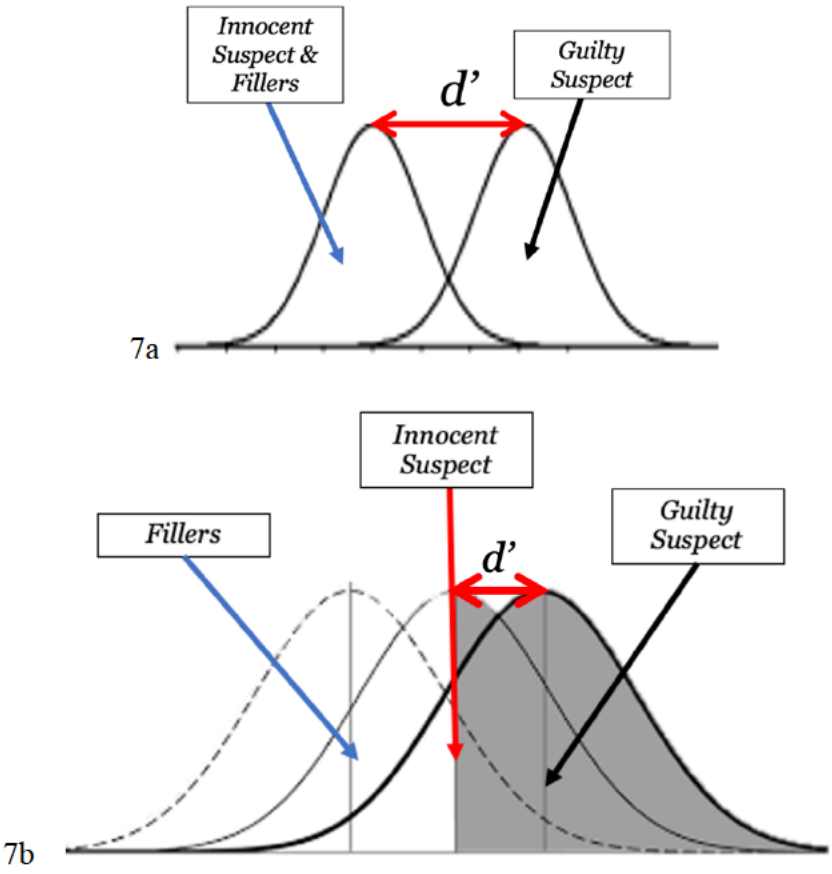
However, the question remains as to how replication lineups should be constructed to optimise witness performance and increase perpetrator identifications in target present lineups, protect innocent suspects from misidentification in target absent conditions, and improve witness's ability to discriminate between innocent and guilty suspects). To consider which replication method (high or low similarity) would result in superior witness performance, signal detection theory (Wixted & Mickes, 2014) and signal detection models, namely diagnostic feature detection (Wixted & Mickes, 2014) and the feature matching model (Colloff et al., 2021) can be used to make predictions about the impact of lineup construction methods on witness identification performance.

According to a signal detection theory interpretation of a lineup task, each lineup member generates a memory signal within the witnesses' memory which they use to make an identification decision (Wixted & Mickes, 2014). The memory signals of faces in lineups can be represented as three memory strength distributions for the guilty suspect, innocent suspect, and fillers. Fair lineups in which the innocent suspect does not stand out will result in two memory distributions, one for the guilty suspect, and one for the innocent suspect and fillers,

because the innocent suspect is no more similar to the guilty suspect than the other lineup members (Wixted & Mickes, 2014; see figure 7a below). However, in unfair lineups, where the suspect stands out, there will be three memory distributions, one for the fillers, innocent suspect, and guilty suspect (see figure 7b). For both fair and unfair lineups, the measure of interest has been the degree of overlap between the guilty and innocent distributions (innocent-guilty suspect discriminability), whereby more overlap would indicate that witnesses are less able to differentiate between innocent and guilty suspects (Colloff et al., 2018; Wixted & Mickes, 2014).

Figure 7

Two distribution model of innocent suspect and fillers, and guilty suspect in a fair lineup (7a) and a three distribution model of fillers, innocent suspect, and guilty suspect in an unfair lineup (7b).



Feature Matching Model (Colloff et al., 2021)

This model assumes that a face may be defined by its number of features and that each facial feature may have several possible settings (Colloff et al., 2021). For example, the feature of eye colour may have settings of brown, blue, hazel, grey and green. It follows that after witnessing a crime, the witness will have stored in memory, the unique features of the perpetrator's face. When presented with a lineup in which the perpetrator is present, the encoded features of the perpetrator in the witnesses' memory will match those of the perpetrator presented in the lineup. However, when presented with a lineup in which the perpetrator is not present, an innocent suspect and fillers, who are not guilty, will not possess the same matching features as they are unique to the perpetrator. In a description matched lineup, fillers are selected for the lineup on the basis that they match the witness's description, and so some of the perpetrator's features will be shared by the fillers and innocent suspect and these features will be non-diagnostic of guilt (Colloff et al., 2021). However, the perpetrator will possess unique features that are not shared by the fillers or innocent suspect in the lineup, which are diagnostic of guilt and can be relied upon by the witness in making an identification decision. Therefore, lineup conditions which maximise the ability of the witness to focus on facial features that are diagnostic of guilt to make an identification decision, will mean that witnesses are more able to identify the guilty perpetrator or correctly reject the innocent suspect.

Diagnostic Feature Detection Theory (Wixted & Mickes, 2014)

Diagnostic feature detection theory states that when faced with a lineup, ability to discriminate between innocent and guilty suspects is improved when witnesses discount non-diagnostic features. Non-diagnostic features are features or characteristics that are not unique to the perpetrator and are therefore not indicative of guilt. Discounting non-diagnostic features and instead focusing on diagnostic features that are unique to the perpetrator

essentially removes noise from the decision process. Diagnostic feature detection theory can be used to explaining the finding that using fair lineups for suspects with distinctive facial features enhances witness's ability to discriminate innocent from guilty suspects compared to unfair (do nothing) lineups (Colloff et al., 2016, 2017). In previous studies (Colloff et al., 2016, 2017), the distinctive feature is non-diagnostic of guilt because both the guilty and innocent suspect have that same distinctive feature (e.g., the have the exact same tattoo). In unfair lineup conditions, where only the suspect has the distinctive feature, the suspect may stand out as similar to the witness's memory of the perpetrator. Therefore, theoretically, witnesses rely on the non-diagnostic distinctive feature when making their identification decision in the unfair condition, damaging their ability to discriminate between innocent and guilty suspects. However, in fair lineups, the non-diagnostic distinctive feature cannot be used in the decision because all lineup members have either the same distinctive feature (replication) or have it covered up (block or pixelation concealment) and so witness will discount the feature and rely on other features that are more diagnostic of guilt.

It is important to note that an alternative filler siphoning perspective (Smith et al., 2019; Smith et al., 2022) argues that unfair lineups do not impair discriminability. Instead, it is suggested that a fair lineup advantage exists because fair lineups lead to a distribution of choices away from the suspect and onto fillers (Smith et al., 2018). While it is acknowledged that filler siphoning occurs to a greater degree in fair compared to unfair lineups, it has been argued that the filler siphoning theory alone cannot explain the fair lineup advantage (Colloff et al., 2018), because filler siphoning does not have a mechanism to a priori predict a discriminability difference across fair and unfair lineup conditions. On the other hand, it has been argued that in fair lineups, filler siphoning predicts an increase in the ability of the investigator to use eyewitness evidence to discriminate between innocent and guilty suspects (Smith et al., 2020; Smith et al., 2022). Furthermore, the ongoing academic debate

surrounding the mechanisms that impact witness performance on fair and unfair lineups highlights the need for further research to test if signal detection-based models are able to make predictions that are confirmed by data and, ultimately, can help to explain the mechanisms of eyewitness memory. Here, this chapter tests predictions made by feature matching model and diagnostic feature detection theory for constructing replication lineups for suspects with distinctive features.

Present Experiment

This experiment investigated whether ability to discriminate between innocent and guilty suspects is improved in lineups for suspects with distinctive facial features, when fillers have a similar but non-identical distinctive feature (low similarity replication), compared to when fillers have a very similar feature (high similarity replication) or when fillers have no distinctive features (do nothing).

Before predictions made by theory are considered, it is important to note a key methodological difference between the current experiment and the previous distinctive feature research (e.g., Colloff et al., 2016; 2017). In previous research by Colloff et al, the innocent suspect had the same distinctive feature as the guilty suspect. This enabled a test of the diagnostic feature detection theory as the feature was non-diagnostic of guilt because it was shared by both innocent and guilty suspects. This also reflected a possible scenario in real life, where the innocent suspect is highly similar to the perpetrator due to being selected because they resemble crime scene footage or a composite sketch of the perpetrator (Quigley-McBride & Wells, 2021; Wells & Penrod, 2011; Wixted & Wells, 2017). However, in practice, the true guilt or innocence of a suspect is not known, and so it is not known how often innocent suspects are presented in a lineup due having a highly similar distinctive feature to the true perpetrator of a crime. Therefore, in the present study, a median similarity innocent suspect was selected to represent the average of the range of possible similarities of

an innocent suspect selected in real life. Furthermore a perpetrator's distinctive feature may therefore be a useful cue to the perpetrator's identity (Valentine, 1991). Therefore, in the work presented here, the innocent suspect will have a similar, but non identical distinctive feature as the perpetrator. Specifically, the innocent suspect's feature will fit the general description ('tattoo on the face') of the perpetrators' feature. Note also that all the features on the faces of the fillers in the low similarity lineups, also fit the general description of the feature ('tattoo on the face').

Research Aims and Hypotheses

The feature matching model (Colloff et al., 2021) can be used to consider the impact of high and low similarity replication lineups and do nothing lineups on witness performance. In *target present conditions*, it is predicted that the hit rate will be lower in the high-similarity replication lineups in comparison to the low similarity replication and do-nothing lineups, and lower in the low similarity replication lineups than then do-nothing lineups. This is because the high similarity replication fillers will share the distinctive facial feature that matches the witness's memory, reducing the number of unique features on the perpetrator in the lineup that match the witness's memory of the perpetrator. As such, high similarity replication fillers compete with the witness's memory of the perpetrator, resulting in a decrease in the hit rate.

Low similarity replication fillers will share fewer features that match the witness's memory of the perpetrator because they each have a different distinctive feature, increasing the number of unique features on the perpetrator (i.e., the distinctive feature) that match the witness's memory of the perpetrator. As such, low similarity fillers compete with the witness's memory of the perpetrator to a lesser extent, resulting in an increase in the hit rate. Similarly, fillers with no distinctive features in the do-nothing lineups compete with the witness's memory of the perpetrator to an even lesser extent, resulting in a further increase in

the hit rate. Put another way, there will be increased filler siphoning in the high similarity replication condition, compared to the low similarity replication condition, and in the low similarity replication condition than the do-nothing condition.

In *target absent conditions*, it is predicted that there will be no difference in the false alarm rate between the low similarity replication and high similarity replication conditions, but the false alarm rate will be higher in the do-nothing condition. A key insight of the feature matching model is that varying filler similarity (i.e., low, or high similarity replication) should not change the number of features on the face of the innocent suspect that match the witness's memory of the perpetrator. Therefore, the false alarm rate should remain unchanged across the low and high similarity replication conditions. In the do-nothing condition the innocent suspect is, however, a better match to memory than the fillers on average because he is the only person with a tattoo, resulting in an increase in the false alarm rate.

Over and above the hit rate and false alarm rate predictions, diagnostic feature detection theory makes predictions about ability to discriminate innocent from guilty suspects (Wixted & Mickes, 2014). Diagnostic feature detection theory predicts that the ability to discriminate innocent from guilty suspects will be lower in the high similarity replication lineups than low similarity replication lineups. In high similarity replication conditions, where the exact same feature appears of every face, witnesses should theoretically discount the tattoo from their identification decision, resulting in a decrease in the available diagnostic features from which the witness can discriminate between innocent and guilty. In low similarity replication lineups, the tattoo differs across the lineup members and so witness should theoretically discount the tattoo to a lesser extent, resulting in an increase in the available diagnostic features from which the witness is able to discriminate innocent from guilty.

The discriminability prediction for the do-nothing lineup condition is, however, less clear. Diagnostic feature detection theory (Wixted & Mickes, 2014) can also be used to predict that the ability to discriminate innocent from guilty suspects will be lower in the high similarity replication lineups than do-nothing replication lineups. In do-nothing lineups, the fillers do not have tattoos and so witnesses have available to them the diagnostic feature (i.e., the tattoo) to aid them in their decision, just like in the low similarity condition. This prediction is tentative, however, as it is based on the premise that participants will properly make use of the of the diagnostic information provided by the tattoos. Put another way, because the innocent suspect's tattoo is not an exact match to the perpetrator's (e.g., the innocent suspect has a star tattoo, while the perpetrator has a tribal tattoo), the prediction assumes that participants will interrogate the shape and design of the tattoo and only make an identification of the suspect if the tattoo exactly matches memory. If, conversely, participants use the mere existence of the tattoo (instead of its exact shape) as evidence of a memory match, this will reduce ability to discriminate innocent from guilty suspects because the mere existence of a tattoo is non-diagnostic indicator of guilt (because both the guilty and innocent suspect have a tattoo of some sort).

Method

Design

Details of the experiment and planned analysis were pre-registered (https://osf.io). The research used a 3 (lineup procedure: high similarity replication, low similarity replication, do nothing) x 2 (target: present, absent) between participant's design. The outcome variables were identification accuracy on the lineup task and participant's confidence in their decision, which was measured on an 11-point Likert scale. While there are no standardised priori power analysis methods for eyewitness identification experiments, other experiments using ROC analysis have included between 300 to 500 participants in each condition (i.e., Colloff et al., 2016, 2017). Therefore, the data collection stopping rule was to recruit at least 1800 participants, with approximately 300 in each condition. Previous research using a designated innocent suspect has found ROC analyses was limited when the data were too noisy to conduct meaningful analyses (Colloff et al., 2021). Therefore, the experiment was pre-registered, and the plan was to repeat data collection with another batch of 1,800 participants until the ROCs appeared stable. To determine that the ROCs were stable, the data points were required to generate a smooth curve on the ROC plot. The research was approved was by the University of Birmingham Science, Technology, Engineering and Mathematics Ethical Review Committee. All participants provided informed consent prior to taking part in the experiment.

Participants

The experiment was conducted using the online platform Qualtrics. An opportunity sample of 6062 participants were recruited via Amazon Mechanical Turk who completed the experiment for financial payment in accordance with local norms ($6.50 per 60 minutes). Participants were from the UK and overseas, they were 16 or over and there was no upper age limit. Data of 1147 participants were excluded as they did not complete the experiment,

experienced technical difficulties, were familiar with stimuli or incorrectly answered an attention-check question. This yielded a final sample size of 4915. The experiment collected the demographic data of participants' gender, age, and ethnicity (See Table 3).

Table 4

Demographic information from Mechanical Turk sample (n =4915)

Characteristics	**Sample**
Sex	
Female	2589
Male	2286
Prefer not to say	40
Age (years)	
M	37.32
SD	13.03
Range	16-91
Prefer not to say	85
Race or Ethnicity	
Asian/Indian	944
Black/African	458
Latin/Hispanic	310
Native American	66
Other	100
White/European	3000
Prefer not to say	37

Materials

Videos

The research used a 31s mugging scenario from Colloff et al. (2016). In the mock crime video, a White male in his late 20s is seen talking on a mobile phone. Then a White male perpetrator in his early 20s approaches and instructs the victim to hand over his mobile phone. When the victim refuses, the perpetrator pushes him, snatches the phone, and flees the scene. The perpetrator has a distinctive facial tattoo on his right cheek, and the camera was

approximately two meters away from the perpetrator when the mock-crime occurs in the video.

In a previous study (Colloff, 2016) collected descriptions by 1460 participants of the perpetrator with a distinctive facial tattoo and a different distinctive perpetrator with a black eye. Colloff (2016) found that almost half of the descriptions provided by participants included specific details about the feature i.e., location and shape (*n*=694). And less than 10% of participants (*n*=138) did not include any details of a distinctive feature in their description of the perpetrator. Therefore, it was assumed that the perpetrators' distinctive tattoo was visible when participants viewed the mock crime video.

Lineups

Lineup members were selected using stimuli from Colloff et al. (2016). To create pools of fillers, Colloff et al. (2016) asked 18 participants to watch the mock crime video. Then participants answered 16 questions regarding the perpetrator's physical attributes, i.e., ethnicity, weight, eye colour, gender, hair colour and height. Then Colloff et al. (2016) entered the modal descriptions into the Florida Department of Corrections Inmate Database (http://www.dc.state.fl.us/AppCommon/) to identify 40 male fillers that matched the modal description of the perpetrator. All fillers faced the camera directly and had neutral facial expressions. Colloff et al. (2016) used Adobe Photoshop–CS5® to transform the filler images to grey scale, remove any background colours, and alter the colour of all filler's t-shirts to black. The perpetrator image from Colloff et al.(2016) was used, he was a white male in his early 20s with a distinctive tribal facial tattoo on the right side of his face. Adobe Photoshop–CS5® was used to adapt 39 stimuli from Colloff et al. (2016) into six pools of filler images (see figure 8).

First, 39 filler images were edited for use in the low similarity replication lineups, and this was completed using findings from a previous study by Colloff (2016). In the study

by Colloff (2016), participants viewed a mock crime video with either a perpetrator with a black eye or a tribal tattoo. The mock crime video of the perpetrator with the tribal tattoo was the same as the video used in this experiment. Then, participants were given two minutes to type a description of the perpetrator in the video. Specifically participants were instructed "Unusual or distinctive features are particularly useful for the police. So please try and describe any unusual or distinctive features in as much detail as possible." Colloff (2016) found that most participants described the distinctive feature correctly and detailed the location of the feature (*n*=2377). Of the remaining participants descriptions, it was more often that the description provided contained less details of the distinctive feature i.e., described something to do with the feature (*n*=2336), when compared to those that provided more details of the distinctive feature i.e., specific location and in detail (*n*=518). The remaining participants either did not describe the distinctive feature, did not complete the task, or did not write a description (*n*=351). Colloff (2016) concluded that overall, most participants did not provide detailed descriptions of the distinctive feature, even when instructed to provide as much detail as possible. Therefore, the present study utilized a general witness description of a 'tattoo on the face' to represent the presence of a distinctive facial feature of a tattoo and the location of the face. Using the general witness description, similar but not identical tattoos were digitally edited onto each of the filler faces. The tattoos varied in design, size, and shape. The location of the tattoos was varied so that 18 fillers had a tattoo on the left side of the face and the remaining fillers had a tattoo positioned on the right side of the face. This was to ensure that the perpetrator did not stand out as the only person with a tattoo on the right side of his face (see figure 8c and 8d).

The experiment intended for target absent lineups to be matched on appearance to a designated innocent suspect. Therefore, to select the innocent suspect, a pilot experiment was conducted to collect ratings of the similarity of each filler from the low similarity replication

pool to the guilty suspect. The low similarity replication filler pool was created using images of 39 males that matched the modal description of the perpetrator. This filler pool was created in a previous study by Colloff et al. (2016). Similar but not identical facial tattoos were edited onto the 39 faces, that matched the general witness description of 'tattoo on the face'. The pilot study was completed by 51 participants on Qualtrics. All participants were required to rate similarity of the 39 perpetrator and filler pairs, using a 7-point Likert scale (where 1= not similar and 7=highly similar). The pilot study was conducted after similar but non-identical tattoos had been digitally edited onto the faces of the fillers. Therefore, the perpetrator had a tribal tattoo on his right cheek and the fillers all had similar but not identical facial tattoos that matched the general witness description of the perpetrator's tattoo. In line with previous research (Colloff et al., 2021) the face with the median similarity rating (M=2.22, SD=1.43) was selected to be the innocent suspect. Therefore, the term median similarity innocent suspect is in the context of description matched fillers, who were selected on the basis of their modal description to the perpetrator (without a tattoo) and then a general description matched facial tattoo was digitally added to the faces of the fillers and the filler who was rated a median similarity (with a distinctive facial feature) was selected as the designated innocent suspect. Again, to reiterate, this meant that in our experiment, the innocent suspect had a distinctive feature that matched the general witness description of the perpetrator's distinctive feature (i.e., tattoo on the face) but was not identical to the perpetrator's distinctive feature. The designated innocent suspect had a star tattoo on his left forehead (see figure 8). The designated innocent suspect was removed from the filler pool to prevent the face appearing twice as an innocent suspect and filler. This resulted in 38 filler images, used in each of the 6 filler pools, described next.

Do nothing fillers. The stimuli for the target present and target absent do nothing lineup conditions did not have a distinctive facial feature (see figure 8e and 8f).

High similarity replication fillers. The stimuli for the high similarity target present filler pool were from Colloff et al. (2016) replication condition. Fillers had the same distinctive tribal tattoo as the perpetrator on the right cheek. For the high similarity target absent filler pool, the star shaped facial tattoo of the designated innocent suspect was replicated across the fillers (see figure 8a and 8b).

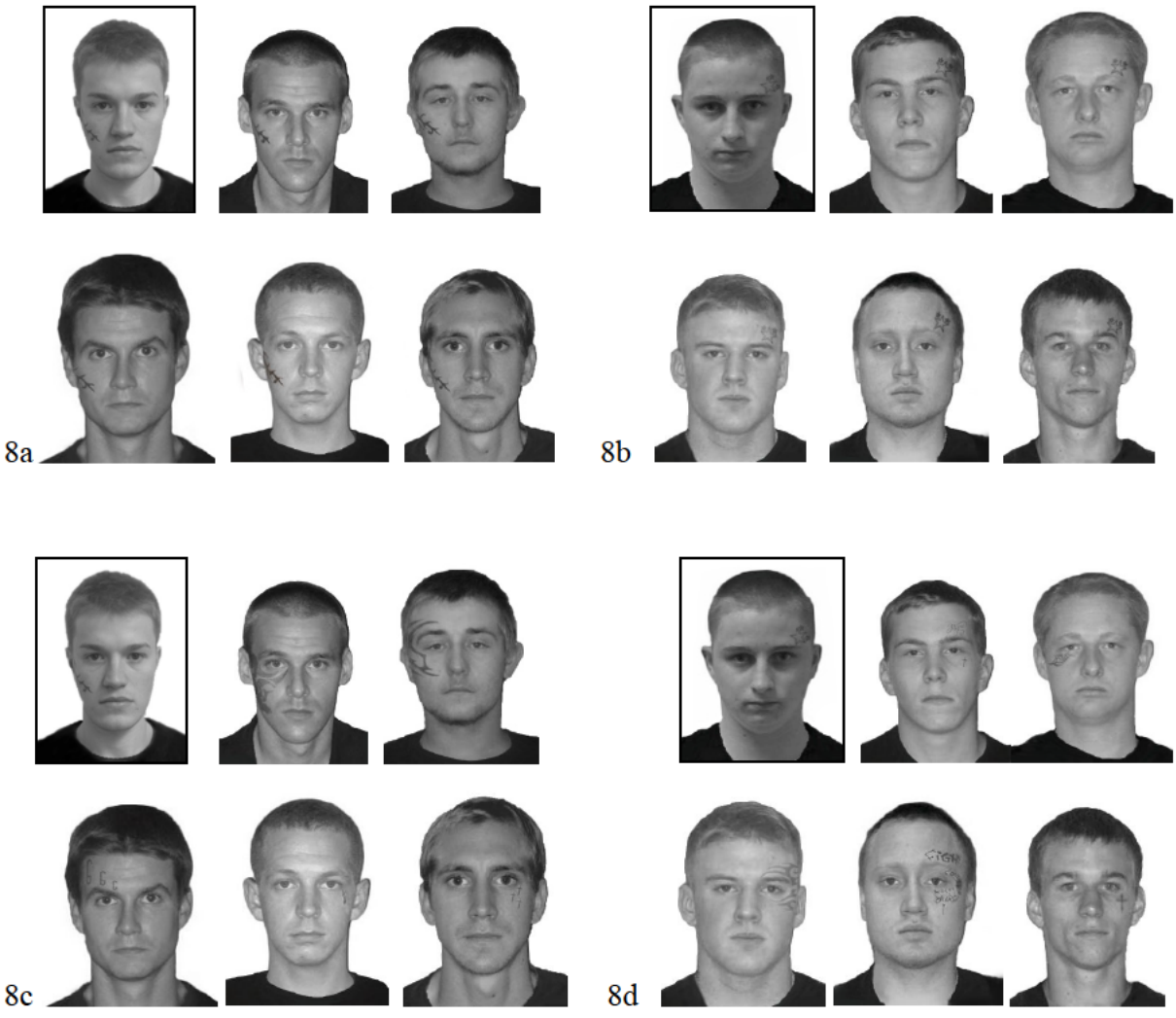
Low similarity replication fillers. As described above, for the stimuli for the target present and target absent low similarity filler pool, similar but not identical tattoos were added that matched witness's description of perpetrator's tattoo onto each of the filler faces (see figure 8c and 8d). Participants rated the similarity of the low similarity replication fillers to the guilty perpetrator using a 7-point Likert scale in a pilot study (described above) whereby similarity ranged from M=1.53 (*SD*=0.67) to M=3.12 (*SD*=1.62).

Pilot testing

Colloff et al. (2016) pilot tested high similarity target present stimuli to ensure that participants were not able to distinguish between digitally edited tattoos on filler faces and the facial tattoo that had not been digitally added onto the perpetrator. For the new low similarity target present stimuli edited for this experiment, a pilot experiment was conducted to test whether participants were not able to distinguish between digitally edited tattoos on filler faces and the facial tattoo that had not been digitally added onto a perpetrator. A total of 45 participants completed the experiment using Qualtrics. Data from 6 participants were excluded because the participants did not complete the full task. All participants were shown a six-person simultaneous lineup containing the perpetrator and five fillers from the low similarity filler pool. Participants were instructed to select which image had not been digitally altered. The results of the pilot confirmed that tattoos on filler faces were realistic and could be used in the experiment, as the perpetrator with the real tattoo was selected 13% of the time, which is not higher that the 17% expected by chance.

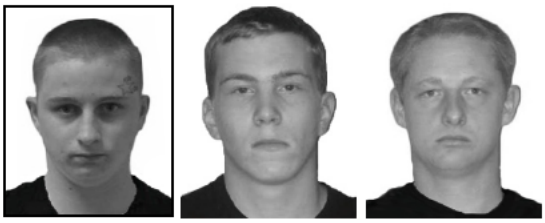
Figure 8

Examples of a target present high similarity replication lineup (8a); target absent high similarity replication lineup (8b); target present low similarity replication lineup (8c); target absent low similarity replication lineup (8d); target present do-nothing lineup (8e); target absent do-nothing lineup (8f).



8a 8b

8c 8d



8e 

8f 

Procedure

Participants watched the mugging mock crime video and were instructed to pay close attention because they would be asked questions about it later. Then participants completed a two-minute anagram task. Participants were then given the following instructions: "Next, you will see 6 faces in a lineup. The perpetrator from the video may or may not be present. If you see the perpetrator from the video, please select him. If you do not see the perpetrator from the video, please select the "Not Present" option." Participants were randomly allocated to one of six experimental conditions (low similarity replication target present, low similarity replication target absent, high similarity replication target present, high similarity replication target absent, do nothing target present, do nothing target absent). All participants viewed a simultaneous lineup of two rows of three photos and a not present option. The target-present lineups contained the perpetrator and 5 fillers randomly generated from the appropriate pool of fillers (low similarity replication, high similarity replication or do nothing). The target absent lineups contained the designated innocent suspect and 5 fillers randomly generated from the appropriate pool of fillers (low similarity replication, high similarity replication or do nothing). The position of the perpetrator or innocent suspect was randomly allocated for each participant.

After making an identification response by clicking on the person who they believed to be the perpetrator or selecting "Not Present", participants were asked to rate their confidence in their decision using a 11-point Likert scale (0% = guessing to 100% = completely certain). Following this, participants were asked demographics questions, if they had experienced any technical difficulties, or if they had seen the video before. Finally, participants were asked what happened in the video to check if they had attended to the task. The experiment was around six minutes in duration. All participants were thanked and debriefed following completion of the experiment.

Results

Participants' identification responses were examined, and ROC analysis was conducted to examine how identification performance was influenced by low and high similarity replication lineups compared to do nothing lineups. For each lineup condition, the proportion of suspect identifications (guilty or innocent), foil identifications and lineup rejections (when participants selected the 'Not Present' option) was calculated and is presented in Table 5.

Table 5

Proportion (and number) of lineup identification responses in high similarity replication, low similarity replication, and do-nothing target present and target absent lineups.

Condition	**High Similarity Replication**	**Low Similarity Replication**	**Do-nothing**
Target Present	*n = 817*	*n = 814*	*n = 815*
Guilty Suspect	.55 *(449)*	.74 *(602)*	.84 *(683)*
Foil	.27 *(219)*	.13 *(103)*	.08 *(65)*
Incorrect Rejection	.18 *(149)*	.13 *(109)*	.08 *(67)*
Target Absent	*n = 834*	*n = 807*	*n = 828*

Innocent Suspect	.02 *(14)*	.01 *(12)*	.07 *(58)*
Foil	.27 *(229)*	.36 *(289)*	.27 *(221)*
Correct Rejection	.71 *(591)*	.63 *(506)*	.66 *(549)*

The rate at which participants identified the guilty suspect (HR) and the innocent suspect (FAR) was compared using z-tests to find out if changes in the HR and FAR across lineup conditions were statistically significant. As predicted by the feature matching model (Colloff et al., 2021), the hit rate was lower in the high similarity replication lineups compared to the low similarity replication lineups ($z = 8.04$, $p < .001$) and do-nothing lineups ($z = 12.73$, $p < .001$). In line with predictions, the hit rate was also lower in the low similarity replication lineup than in do nothing lineups ($z = 4.86$, $p < .001$). The reduction in the hit rate in the high similarity condition compared to the low similarity condition and the low similarity condition compared to the do-nothing condition, was accompanied with an increase in the filler identification rates.

As predicted by the feature matching model (Colloff et al., 2021), there was no difference in the false alarm rate between the low and high similarity replication lineups ($z = 0.32$, $p = .75$). In line with predictions, the false alarm rate was higher in the do-nothing lineups compared to the low similarity replication lineups ($z = 5.50$, $p < .001$) and high similarity replication lineups ($z = 5.30$, $p < .001$).

Discriminability

Next, to test if there was a difference in ability to discriminate innocent from guilty suspects across the lineup conditions, Receiver Operator Characteristic (ROC) analysis was conducted. ROC curves were constructed for each of the lineup conditions; high similarity replication, low similarity replication and do-nothing (see Figure 2).

To create the ROC curves, participants confidence ratings, Hit Rates (HR) and False Alarm Rates (FAR) were used. Participants rated their confidence in their identification

decision on an 11-point Likert scale from 0% to 100% in intervals of 10 (100, 90, 80, 70 etc). The HR was the number of times a guilty suspect was correctly identified in a lineup in which the guilty suspect was present (target present), divided by the total number of target-present lineups. The FAR was the number of times that an innocent suspect was identified in a lineup in which the innocent suspect was present (target absent), divided by the total number of target-absent lineups. For each lineup condition (high similarity replication, low similarity replication and do-nothing), the ROC curve plots the HR and FAR over decreasing levels of confidence. This means that the left-most points of the ROC curve represents the HR and FAR at the highest rating of confidence (i.e., when participants were 100% certain of their identification decision). Then, the second left-most point on the ROC curve depicts the cumulative HR and FAR at the two highest levels of confidence (e.g. 90% and 100% certain). The ROC curve continues to show HR and FAR pairs at cumulative levels of confidence until the right-most point which shows the HR and FAR of all participants that made a suspect identification (see Mickes et al., 2012).

To find out if differences in discriminability were statistically significant in the three lineup conditions, partial area under the curve (*p*AUC) values were calculated using the statistical package pROC (Robin et al., 2011) with RStudio (RStudio Team, 2021) and the R software environment (Version 3.2.0; R Development Core Team, 2021). Three pairwise comparisons were completed using *D*, which is the difference of the two *p*AUCs divided by the standard deviation of the difference estimated by bootstrapping (using the pROC; Robin et al. 2011). Specificity (1 – FAR) when calculating the *p*AUC was set using the smallest FAR range, which was .015 (from low similarity replication condition) and so the specificity was set to .985.

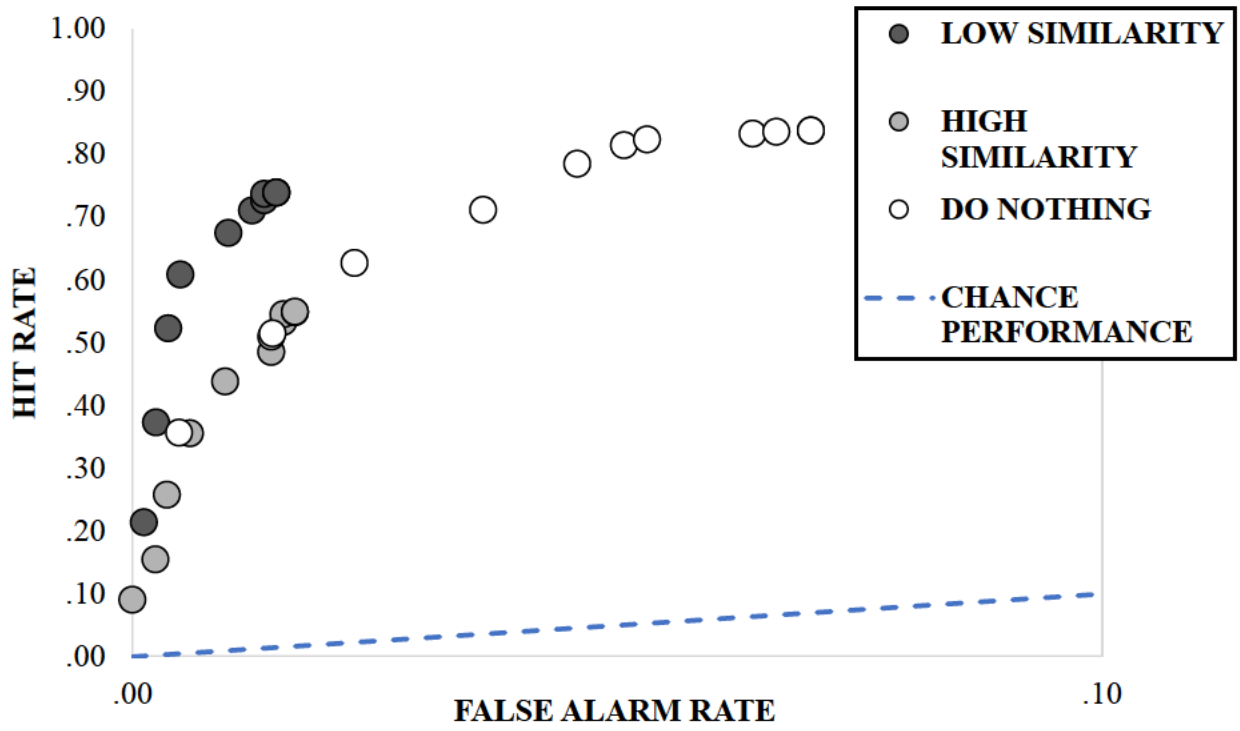
The pAUC for the low similarity replication lineup was larger (*pAUC* = 0.009, 95% CI: 0.006 – 0.010), than the pAUC for high similarity replication lineups (*pAUC* = 0.005,

95% CI: 0.004 – 0.007), and this difference was statistically significant (D = 2.456, p = .01). This indicates that, as predicted by diagnostic feature detection theory (Wixted & Mickes, 2014), ability to discriminate innocent from guilty suspects was better in low similarity replication lineups, in which the distinctive feature was varied across lineup members, compared to high similarity replication lineups, in which all lineup members had the same distinctive facial feature.

The pAUC for the low similarity replication lineups was larger than the pAUC for do nothing lineups ($pAUC$ = 0.005, 95% CI: 0.004 – 0.007), and this difference was statistically significant (D = 2.324, p = .02). This indicates that, ability to discriminate innocent from guilty suspects was better in low similarity replication lineups, in which the distinctive feature was varied across lineup members, compared to do nothing lineups, in which the suspect (innocent or guilty) stood out as being the only person with a distinctive facial tattoo. There was no significant difference in the pAUC for the do-nothing lineups compared to the pAUC of the high similarity lineups, D = 0.079, p = .94. This indicates that, contrary to our predictions, ability to discriminate innocent from guilty suspects was not better in do-nothing lineups compared to high similarity lineups. That is, when presented with a lineup in which the suspect (innocent or guilty) stood out as being the only person with a distinctive facial tattoo, participants were no more able to discriminate between innocent and guilty suspects than when they were faced with a lineup in which all lineup members had the same distinctive facial feature.

Figure 9

Receiver operating characteristic (ROC) curves of the low similarity replication, high similarity replication and do-nothing lineups. Chance performance is shown by the dashed line.



Discussion

This experiment investigated whether ability to discriminate between innocent and guilty suspects was improved in lineups for suspects with distinctive facial features when fillers have a similar but non-identical distinctive feature (low similarity replication), compared to when fillers have a very similar feature (high similarity replication) or when fillers have no distinctive features (do nothing). In doing so, predictions of the feature matching model (Colloff et al., 2021) and diagnostic feature detection theory (Wixted & Mickes, 2014) were tested.

Feature matching model predictions

According to the feature matching model (Colloff et al., 2021), in *target present conditions*, it was predicted that the hit rate would be lower in the high similarity replication lineups in comparison to the low similarity replication and do-nothing lineups, and lower in the low similarity replication lineups than then do-nothing lineups. Put another way, there would be increased filler siphoning in the high similarity replication condition, compared to the low similarity replication condition, and in the low similarity replication condition than the do-nothing condition.

Our results revealed that theoretical predictions (Colloff et al., 2021) were supported. The hit rate was significantly lower in the high similarity replication lineups compared to the low similarity replication lineups and do-nothing lineups. This indicates that, as predicted by the feature matching model (Colloff et al., 2021), in high similarity replication lineups, participants identified the guilty perpetrator significantly less often than in low similarity replication lineups and do-nothing lineups. That is, when presented with a lineup in which all lineup members had the same distinctive facial feature, participants identified the guilty perpetrator less often than when presented with a lineup in which the distinctive feature was varied across lineup members, or in which the guilty perpetrator was the only person with the distinctive facial feature.

Furthermore, in line with predictions, the hit rate was also significantly lower in the low similarity replication lineup than in do-nothing lineups. This indicates that, as predicted by the feature matching model (Colloff et al., 2021), in low similarity replication lineups, participants identified the guilty perpetrator significantly less often than in do-nothing lineups. That is, when presented with a lineup in which the distinctive feature was varied across lineup members, participants identified the guilty perpetrator less often than when

presented with a lineup in which the guilty perpetrator was the only person with the distinctive facial feature.

According to the feature matching model (Colloff et al., 2021), in *target absent conditions*, there should be no difference in the false alarm rate between the low similarity replication and high similarity replication conditions, but the false alarm rate would be higher in the do-nothing condition. Our results revealed that, as predicted by the feature matching model (Colloff et al., 2021), there was no significant difference in the false alarm rate between the low and high similarity replication lineups. This indicates that, in low similarity replication lineups and high similarity replication lineups, there was no difference in the identification rate of the innocent suspect. That is, when presented with a lineup in which the innocent suspect had a different distinctive facial feature to the guilty perpetrator, there was no difference in the identification of the innocent suspect in lineups where all members had the same distinctive facial feature, compared to lineups in which all lineup members had varied distinctive facial features.

Also in line with predictions (Colloff et al., 2021), the false alarm rate was higher in the do-nothing lineups compared to the low similarity replication lineups and high similarity replication lineups. This indicates that, participants identified the innocent suspect significantly more often in do-nothing lineups compared to low similarity replication lineups and high similarity replication lineups. Again, the innocent suspect within these target absent lineups had a distinctive facial feature (e.g., star) that was different to that of the guilty perpetrator (e.g., tribal). That is, when presented with a lineup in which the innocent suspect was the only line up member with a distinctive tattoo, even though that tattoo did not match that of the guilty perpetrator, participants identified the innocent suspect more often than when faced with lineups in which lineup members had the same distinctive facial feature or a varied distinctive facial feature.

Diagnostic feature detection predictions

According to diagnostic feature detection theory (Wixted & Mickes, 2014), it was predicted that ability to discriminate innocent from guilty suspects would be better in low similarity replication lineups compared to high similarity replication lineups. Consistent with predictions, our results indicated that ability to discriminate innocent from guilty suspects was better in low similarity replication lineups, in which the distinctive feature was varied across lineup members, compared to high similarity replication lineups, in which all lineup members had the same distinctive facial feature. Moreover, it was found that ability to discriminate innocent from guilty suspects was better in low similarity replication lineups, in which the distinctive feature was varied across lineup members, compared to do nothing lineups, in which the suspect (innocent or guilty) stood out as being the only person with a distinctive facial tattoo.

According to diagnostic feature detection theory (Wixted & Mickes, 2014), it was also hypothesised that ability to discriminate innocent from guilty suspects would be better in do nothing lineups compared to high similarity replication lineups. Contrary to our predictions, our results indicated that ability to discriminate innocent from guilty suspects was not better in do-nothing lineups compared to high similarity replication lineups. That is, when presented with a lineup in which the suspect (innocent or guilty) stood out as being the only person with a distinctive facial tattoo, participants were not better able to discriminate between innocent and guilty suspects than when they were faced with a lineup in which all lineup members had the same distinctive facial feature. But why were predictions not supported in our results? explanations for this finding are considered next.

Do-nothing lineups

Using theoretical models (Colloff et al., 2021; Wixted & Mickes, 2014), patterns of suspect identification (hit rate and false alarm rate) and ability to discriminate between

innocent and guilty suspects in do-nothing lineups were predicted. In do-nothing lineups, only the suspect (guilty or innocent) had a distinctive facial feature, and the other fillers did not. In target present conditions, the guilty suspect had a tribal tattoo. And in target absent conditions, the innocent suspect had a star tattoo. Specifically, the innocent suspect's tattoo fit the general description of the perpetrators' distinctive feature (i.e., tattoo on the face). This is of particular importance, as in previous research, experiments have used an innocent suspect with the same distinctive feature (i.e., tribal tattoo) as the guilty suspect (Colloff et al., 2016, 2017). Furthermore, previous research enabled a test of the diagnostic feature detection theory (as the feature was non-diagnostic of guilt), and also reflected a possible scenario in real life. That is, where the innocent suspect is highly similar to the perpetrator, and this is due to the innocent suspect being selected because they resemble crime scene footage or a composite sketch of the perpetrator (Quigley-McBride & Wells, 2021; Wells & Penrod, 2011; Wixted & Wells, 2017). However, in practice, the true guilt or innocence of a suspect is not known, and so it is not known how often innocent suspects are presented in a lineup due having a highly similar distinctive feature to the true perpetrator of a crime. Therefore, in the present study, a median similarity innocent suspect was selected to represent the average of the range of possible similarities of an innocent suspect selected in real life. Therefore, the similar but non-identical distinctive feature of the innocent suspect presented in our study may be a useful cue to the perpetrator's identity (Valentine, 1991).

On that basis, it was predicted that in do-nothing lineups, fillers do not have tattoos and so witnesses have available to them the diagnostic feature (i.e., the tattoo) to aid them in their decision. Put another way, because the innocent suspect's tattoo is not an exact match to the perpetrator's (e.g., the innocent suspect has a star tattoo, while the perpetrator has a tribal tattoo), the prediction assumes that participants will interrogate the shape and design of

the tattoo and only make an identification of the suspect if the tattoo exactly matches memory.

However, the pattern of results in the present study suggests that participants did not use the shape and design of the tattoo to inform their lineup identification decision. This is evidenced by the increased hit rate and false alarm rate in the do nothing condition compared to the low similarity replication and high similarity replication conditions. This suggests that participants gloss over the characteristics of the distinctive feature and the mere existence of the tattoo (instead of its exact shape) was used as evidence of a memory match. And this resulted in an increase in identifications in both target present and target absent lineups. As a result, ability to discriminate innocent from guilty suspects is reduced because the mere existence of a tattoo is a non-diagnostic indicator of guilt (because both the guilty and innocent suspect have a facial tattoo). This is further supported by the finding that ability to discriminate innocent from guilty suspects was better in low similarity replication lineups. That is, in the low similarity replication lineups, participants are required to scrutinise the shape and the design of the facial tattoo to determine which is a match to memory. And this results in an improved ability to discriminate between innocence and guilty.

Ability to discriminate between innocent and guilty suspects was the same in the do-nothing and high similarity replication lineups. It is argued that in the do-nothing conditions, participants picked the suspect due to the mere existence of a facial tattoo, regardless of their guilt or innocence, harming their discriminability compared to the low similarity replication condition. Yet, in the high similarity replication condition, all lineup members had the exact same distinctive feature. That is, in target present conditions, the guilty suspect and fillers had a tribal tattoo. And in target absent conditions, the innocent suspect and fillers had a star tattoo. So it is not possible that the mere existence of a tattoo can explain the pattern of lineup identification response in high similarity replication lineups, because all lineup members had

the same tattoo. Instead, in high similarity replication conditions, participants will discount the facial tattoo, as it is non-diagnostic of guilt (Colloff et al., 2021; Wixted & Mickes, 2014). That is, because the facial tattoo is shared by all lineup members, it is not indicative of guilt and will not aid participants in their identification decision (Colloff et al., 2021). As such, there are fewer unique features of the guilty suspect to aid identification, because the lineup members are highly similar (see Chapter 2). The result is a decrease in ability to discriminate between innocent and guilty in high similarity replication lineups compared to low similarity replication lineups.

Therefore, the finding that both do-nothing and high similarity replication conditions yielded the same level of discriminability was likely to be a mere coincidence. That is, it appears that there were separate processes underlying the seemingly equivalent performance. In do-nothing conditions, that was glossing over the distinctive feature characteristics and using the mere existence of a facial tattoo as evidence of a memory match, regardless of the guilt or innocence of the suspect. And in high similarity replication conditions, that was discounting of a useful feature (tattoo) because it was shared by all lineup members. As a result, both processes result in an overall decrease in the ability to discriminate between innocent and guilty suspects compared to the low similarity replication condition. It is argued that scrutinising the facial tattoo in low similarity replication conditions results in improved ability to discriminate between innocent and guilty suspects.

Implications

The present study has tested and supported theoretical predictions (Colloff et al., 2021; Wixted & Mickes, 2014) of optimal lineup construction methods for distinctive suspects. It has been demonstrated that constructing a lineup in which all lineup members have a similar but not identical distinctive feature (low similarity replication), improves ability to discriminate innocent from guilty, while also protecting the innocent suspect from

misidentification. Moreover, theoretical accounts of lineup performance (Colloff et al., 2021; Wixted & Mickes, 2014) have been developed further by highlighting the different process involved in lineup conditions. That is, feature discounting in high similarity replication lineups, feature scrutinising in low similarity replication lineups, and feature glossing over in do-nothing lineups. And it is argued that these different processes are evidenced by the findings of the present study. Future research should replicate our findings and establish if the observed patterns of results are consistent across multiple samples and different lineups.

Furthermore, the present study contributes to the existing literature on optimal lineup construction. For example, our findings support theoretically similar research by Colloff et al. (2021). Namely, that witness identification performance is enhanced when lineups are constructed by selecting fillers who are dissimilar to the perpetrator, but within the constraints of the witness description (i.e., low similarity replication lineups). On the other hand, our findings do contradict conceptually similar results. Previous meta-analyses and reviews that have argued lineup members should be moderately to highly similar to protect the innocent suspect from misidentification (Fitzgerald et al., 2013, 2015). But in the present study, the highly similar lineups (high similarity replication) did not enhance ability to discriminate innocent from guilty suspects. However, the discrepancy in results of the present study and previous reviews (Fitzgerald et al., 2013, 2015) is likely to exist because of methodological inconsistencies within the existing literature that have made it difficult to reliably interpret previous studies of filler similarity (see Chapter 2).

Additionally, our results have implications for how lineups are constructed in practice. Both in the U.S and the UK, there is variability in the lineup construction methods used for distinctive suspects (Police and Criminal Evidence Act 1984, Code D, 2011; Technical Working Group for Eyewitness Evidence, 1999; Wolgater et al., 2004). In the UK, one option is to replicate the distinctive facial feature across the other lineup members.

However this method is currently not standardised within practice. For example, a replication lineup for a distinctive suspect with a tattoo on the right cheek could include fillers with the exact same tattoo on the right cheek (high similarity replication), or the tattoo style and location could be varied across fillers (low similarity replication; see figure 6). The present study provides the first empirical test of high and low similarity replication methods for distinctive suspects. It is argued that, when constructing lineups for distinctive suspects, replicating a similar but non identical distinctive feature across other lineup members will result in an improved ability to discriminate between innocent and guilty suspects. And it is recommended that more empirical studies are conducted to establish a robust evidence base for the low similarity replication advantage, so that this can, in the future, be communicated to policy makers to inform police practice. Thus, enabling witnesses to better identify guilty suspects and reject the innocent suspects in real life.

However, it is important to note the constraints on generality (Simons et al., 2017) of the present research. Firstly, the encoding conditions of the experiment involved only one mock-crime video depicting the perpetrator with the tribal tattoo at approximately two metres distance. On the basis of a previous study that collected witness descriptions of the same perpetrator with a distinctive tribal tattoo (Colloff, 2016) it was assumed that the distinctive tattoo was visible when participants watched the mock-crime video. However, participants in the present experiment were not asked to provide a description of the distinctive suspect and so the extent to which they encoded the details of the distinctive facial feature is not known. Furthermore, the use of only one mock-crime video means that the conclusions of the present research are limited to situations where there are similar encoding conditions. That is, there may be a different pattern of results when encoding conditions are different. For example if there was a closer view of the perpetrators' distinctive facial feature, participants may provide a more detailed description of the feature and so a low similarity replication lineup

may not be possible in practice. That is, when a witness has provided a detailed description of the distinctive feature, which could be due to improved encoding conditions (i.e., a better view of the distinctive feature) then the possible variation of the distinctive features replicated across other lineup members may be limited. For example, if the witness description states 'tattoo of a rose on the face' then a lineup member with a different shaped tattoo (i.e., a star tattoo) would not match the witness description. Instead, a low similarity replication lineup would need to ensure that all lineup members' distinctive feature matched the description of 'tattoo of a rose on the face' and the low similarity replication would be of a facial tattoo of a rose in different shapes, styles and sizes across lineup members. Furthermore, in practice this may mean that law enforcement are unlikely to be investigating a suspect with a star tattoo on the face (such as that of the median similarity innocent suspect in the present study) if the witness has given a detailed description of the perpetrator having a rose tattoo on the face.

Therefore, the findings of the present study are limited to circumstances whereby the witness has provided a general description of the suspect's distinctive facial feature (i.e., 'tattoo on the face'). As it is only when a general description has been provided (i.e., 'tattoo on the face'), that it is possible to construct a low similarity replication lineup such as that in the present experiment, that matches the witnesses' description of the perpetrator. And it is only when a general description has been provided (i.e., 'tattoo on the face'), that it is possible for an innocent suspect to be selected on the basis of having a facial tattoo (which does not resemble that of the real perpetrator), while still matching the witness description of the perpetrator. Therefore, the use of a median similarity innocent suspect may also be limited to circumstances when the witness has not provided a detailed description of the perpetrators' distinctive feature. And in circumstances where a detailed description of the perpetrators' distinctive feature has been provided (i.e., 'rose tattoo on the face'), then it is

likely that the innocent suspect will also be highly similar to the perpetrator as they will possess a highly similar distinctive feature in order to match the description of the perpetrator.

Conclusion

Overall, the present experiment supports theoretical predictions on constructing lineups for distinctive suspects (Colloff et al., 2021; Wixted & Mickes, 2014). And the theory has been developed further by highlighting the different processes (i.e., feature discounting, scrutinising, and glossing) in high similarity replication, low similarity replication, and do-nothing lineups. It is recommended that, when constructing lineups for distinctive suspects, law enforcement personal should replicate a similar, but non-identical distinctive feature across the lineup members. In doing so, this will improve witness ability to discriminate between innocent and guilty, without putting innocent suspects at increased risk of misidentification. However, the constraints on generality should be noted (Simons et al., 2017), as the present results are limited to conditions in which the witness has provided a general description of the distinctive facial feature (i.e., ‘tattoo on the face’). Nevertheless, if our results are replicated by other researchers, these findings could inform recommendations for practice.

CHAPTER 4 : A PSYCHOMETRIC CRITIQUE OF THE BENTON FACIAL RECOGNITION TEST

Abstract

The present chapter considers the Benton Facial Recognition Test (BFRT), which is used to assess clinical impairment in unfamiliar facial recognition ability. Four versions of the BFRT are presented: short form, long form, computerised (BFRTc) and revised (BFRTr). This chapter examines the psychometric properties of all versions of the BFRT, including ability to measure the construct of unfamiliar facial recognition (i.e., validity) and to obtain a consistent score when the test is repeated, (i.e., reliability). Overall, it was reported that all versions of the BFRT did measure the construct of unfamiliar facial recognition ability, and therefore demonstrated validity. Furthermore, the long form BFRT, BFRTc and BFRTr did evidence reliability, but the short form BFRT did not demonstrate acceptable reliability. This chapter also present results of clinical and non-clinical populations' performance on the BFRT to date, and the extent to which the BFRT is able to identify an individual with facial processing deficits when they do indeed have face processing deficits. The results indicate that the sensitivity of BFRT, BFRTc and BFRTr is poor, and research has highlighted that clinical populations (i.e., with impairment in facial recognition ability) can score within the average range. Therefore, it is recommended that there is development of culturally sensitive versions of the BFRT that are appropriate for use in non-Caucasian populations. Finally, this chapter considers the applied use of the BFRT as a measure of individual differences in facial recognition ability within the eyewitness lineup identification paradigm. It was identified that the BFRT does show promise as an indicator of eyewitness lineup identification accuracy, however the relationship between BFRT scores and lineup identification accuracy is impacted by age of participant and lineup type.

Introduction

Often in research, it is assumed that there are no individual differences between witnesses in experimental conditions. However, recognition ability within the general population is diverse (Burton et al., 2010; Darling et al., 2009; Russell et al., 2009; Wang et al., 2012; Wilmer et al., 2012; Woodhead & Baddeley, 1981; Wilmer et al., 2010; Zhu et al., 2010). That is, facial recognition ability ranges from above average (i.e., a 'super-recognizer', Ramon et al., 2019; Russel et al., 2009) to individuals with clinical impairments in facial recognition ability (i.e., 'prosopagnosia', Barton & Carrow, 2016). This means that, within a group of witnesses, some people may be better able to identify a guilty suspect than others. And the identification of a suspect in a lineup is considered to be a task of unfamiliar facial recognition, because the suspect is not known to the witness (Young & Burton, 2017). Furthermore, it is argued that the eyewitness literature has generally overlooked the heterogeneity of unfamiliar face recognition ability (Grabman & Dodson, 2020).

Therefore, a measure of individual facial recognition ability may be useful to consider lineup identification outcomes, such as the accuracy of a suspect identification. Indeed, research has identified a correlation between performance on facial recognition tests and eyewitness identification accuracy (Binderman et al., 2012; Geiselman et al., 2003; Gettleman et al., 2021; Graham et al., 2019; Hosch, 1994; Memon et al., 2003; Morgan et al., 2007; Searcy et al., 1999; Searcy et al., 2001). So, the use of facial recognition tests within research may be beneficial to evaluate if individual differences in facial recognition ability exist and are associated with lineup performance. And if they do exist, to identify if individual differences in facial recognition ability influence the effect of a lineup manipulation on performance in experimental studies. To explore this, this chapter will consider the Benton Facial Recognition Test (BFRT) as a measure of facial recognition ability and an indicator of eyewitness accuracy.

Overview of the BFRT

The BFRT is an assessment of facial recognition (Benton et al., 1994). It was created to identify clinical impairment in facial recognition ability, known as prosopagnosia (Benton & Van Allen, 1968; Barton & Carrow, 2016). In the initial study, Benton and Van Allen (1968) administered the BFRT to clinical and non-clinical samples (15 participants with right hemisphere brain disease, 22 participants with left hemisphere brain disease, and 111 control participants) and found that performance on the BFRT was significantly lower in the clinical sample. Moreover, participants with brain disease in the right hemisphere obtained significantly lower scores on the BFRT than participants with left hemisphere brain disease. Benton and Van Allen (1968) concluded that impairment in the recognition of faces, as measured by the BFRT, is associated with brain disease in the right hemisphere.

Subsequent research has identified differences in facial recognition for familiar and unfamiliar faces. A familiar face is that of a known individual, such as a family member, and an unfamiliar face is that of an unknown individual (Young & Burton. 2017). That is, studies have found that patients who had impairments in familiar face recognition could recognise unfamiliar faces and patients with impairments in unfamiliar face recognition were able to recognise familiar faces (Benton & Van Allen, 1972; De Renzi, 1986; Rondot et al., 1967; Tzavaras et al., 1973). It is agreed within the literature that facial recognition can be impaired in two areas: familiar face recognition and unfamiliar face recognition. Therefore, Benton and colleagues describe the BFRT as a standardised test of unfamiliar facial recognition only (Benton et al., 1994).

Previous Versions of the BFRT

Initially, there were two versions of the BFRT: the long form and the short form (Benton & Van Allen, 1968; Benton et al., 1994; Levin et al., 1975). Both the long form and short form tests are administered with a stimulus booklet, whereby participants are asked

“You see this [man or woman], show me where [he or she] is on this picture”. Then, both the short form and long form BFRT have three parts that involve matching identical photos of Caucasian faces and matching different photos of the same Caucasian face (Benton & Van Allen, 1968).

In part A, participants are required to match identical front-view photographs. A front view photograph of a target face is presented, and participants are requested to identify the target from a six-person simultaneous display of photographs. In all conditions, the target face remains on the screen when the six person array is shown. In both the short form and long form test, there are six trials resulting in a total of six responses. Over the six trials, a total of three male and three female target faces are presented for matching (Benton et al., 1994).

In part B, the task is to match front-view and three-quarter view photographs. Subjects are provided with a front-view photograph of a target face and requested to find the target three times within in each trial of a six-person simultaneous display. In each six-person simultaneous display, all photographs are presented in three-quarter view. In the long form test, there are eight trials resulting in a total of twenty four responses. Over the eight trials, a total of four male and four female target faces are presented for matching. In the short form test, there are four trials resulting in a total of twelve responses. Over the four trials, a total of one male and three female target faces are presented for matching.

Part C requires participants to match front-view photographs in different lighting conditions. A front-view photograph of the target is presented, in which the image was taken under full lighting conditions. The participant is instructed to find the target three times within each trial of a six-person simultaneous display. In each six-person simultaneous display, all photographs are presented in front-view and were taken under altered lighting conditions. In the long form test, there are eight trials resulting in a total of twenty four

responses. Over the eight trials, a total of four male and four female target faces are presented for matching. In the short form test, there are three trials resulting in a total of nine responses. Over the three trials, a total of two male and one female target faces are presented for matching.

Each correct answer results in a score of one and an incorrect answer results in a score of zero. The long form BFRT consists of 54 scoreable responses and requires twenty minutes for administration (Benton & Van Allen, 1968). The short form BFRT consists of 27 scorable responses and requires seven minutes for administration, with a variation of five to fifteen minutes depending on participant ability (Benton et al., 1994; Levin et al., 1975). On the long form test, a score of 25 can be achieved by chance alone and so the range of scores to be considered are between 25 and 54 (Benton et al., 1994). On the short form test, a score of 11 can be achieved by chance alone and so the range of scores to be considered are between 11 and 27 (Benton et al., 1994).

A record sheet is used to record and score performance on the short form and long form BFRT (Benton et al., 1994). In the short form BFRT, the total number of correct responses is calculated, and a conversion table is used (see Table 6) to convert the short form score to an equivalent long form score. Then a corrected long form score is obtained using a corrections table that considers age and years in education (see Table 7). Similarly, in the long form BFRT, the total number of correct responses is calculated, then a corrected long form score is obtained using a corrections table that considers age and years in education (see Table 7). Corrected scores for both the short form and long form BFRT are then interpreted using normative standards (see Table 8).

Table 6

BFRT Short Form to Long Form Score Conversions (Benton et al., 1994).

Short Form	Long Form
27	54
26	52
25	50
24	49
23	47
22	45
21	43
20	41
19	39
18	37
17	36
16	34
15	32
14	30
13	28
12	27
11	25

Table 7

BFRT Score Corrections (Benton et al., 1994).

	Years in Education	
Age (years)	***6-11***	***12+***
16-54	0	0
55-64	3	1
65-74	4	2

Table 8

Facial Recognition Normative Standards (Benton et al., 1994).

Corrected Score	Percentile Rank	Classification
53-54	98+	Very Superior
50-52	88-97	Superior
47-49	72-85	High Average
43-46	33-59	Average
41-42	16-21	Low Average
39-40	8-11	Borderline
37-38	3-6	Defective
<37	1	Severely Defective

New Versions of the BFRT

Since the development of the short and long form BFRT (Benton & Van Allen, 1968; Benton et al., 1994; Levin et al., 1975), there have been two further versions of the BFRT that involve computerised (BFRTc, Rossion & Michel, 2018) and revised materials (BFRTr, Murray et al., 2021). The key difference is that the BFRTc uses the same stimuli from the long form BFRT and the BFRTr uses updated stimuli (Murray et al., 2021).

Similarly to the long form BFRT, the BFRTc and BFRTr include twenty-two trials of selecting a target from conditions of front facing images (part A), varying camera angle (part B) and lighting conditions (part C). Both the BFRTc and BFRTr are administrated digitally, and trials are presented in the same order for each participant. Furthermore, there is no time limit on completion time for each trial and there is an 800ms interval between each of the trials (Murray et al., 2021; Rossion & Michel, 2018). In both the BFRTc and BFRTr, the stimuli are Caucasian faces (Murray et al., 2021; Rossion & Michel, 2018). Responses on the BFRTc and BFRTr are scored in the same way as the long form BFRT, corrected scores are

calculated (see Table 7), and normative data is used (see Table 8) to interpret overall unfamiliar facial recognition ability.

There are three key differences between the BFRTc and the long form BFRT. Firstly, in the BFRTc, target and distractor faces are presented on the same dark background together, unlike the BFRT which presents target and distractor faces on separate panels of the stimulus booklet. Furthermore, the size of the images are slightly larger in the BFRTc (133 x 200 pixels) when compared with the BFRT (129 x 150 pixels). Finally, the BFRTc differs from the original as participants are told to complete the task as fast as possible (Rossion & Michel, 2018).

Furthermore, the BFRTr differs from previous versions, in that it includes male faces only (Murray et al., 2021). The authors report this methodological decision was due to a growing evidence base of an own-gender bias for female face recognition but not for male own-gender face recognition (Herlitz & Lovén, 2013; Lovén et al., 2011; Murray et al., 2021). That is, there is evidence that females demonstrate improved facial recognition for female compared to male faces, whereas males do not display improved facial recognition for male faces. Murray and colleagues report that the use of male only stimuli in the BFRTr is consistent with other neuropsychological tests of facial perception (Duchaine & Nakayama, 2006; Duchaine et al., 2007). Moreover, the key distinction of the BFRTr is that the stimuli are changed to include more varied and naturalistic facial images compared to the original stimuli (Murray et al., 2021). Specifically, the BFRTc includes varied images for each male target, that were taken on different dates within a one year period, in which hairstyle, skin tone, blemishes and lighting varied (Murray et al., 2021).

Other measures

Other commonly used tests of facial recognition include the Cambridge Face Memory Test (CMFT; Duchaine & Nakayama, 2006). Unlike the BFRT, which is a face matching task

(Benton et al., 1994), the CMFT involves recognition of faces from previously viewed learning images. Research has argued that the CMFT is more sensitive than the BFRT in identifying individuals with prosopagnosia (Bowles et al., 2009; Duchaine, & Nakayama, 2004; 2006). Furthermore, additional measures of face perception and face matching include the Cambridge Face Perception Test (CPFT; Duchaine et al., 2007) the Glasgow Face Matching Test (Burton et al., 2010) and the Pairs Matching Test (Bate et al., 2018; Bate et al., 2019). These measures have been criticised for being overly complex and lacking in sensitivity to identify deficits such as prosopagnosia (Bate et al., 2018; Bate et al., 2019; Bowles et al., 2009). In contrast, the BFRT has simple instructions, a relatively fast administration time and it does not have a ceiling effect (i.e., most participants are not able to achieve the highest score), and so it is considered to be a difficult test with interpretable responses (Benton & Van Allen, 1972).

Critical Evaluation of the BFRT

Next, this chapter will evaluate how effectively the BFRT measures unfamiliar facial recognition ability and consider its applicability to eyewitness research. In general evaluation of psychometric measures, a good test should be reliable, valid, and discriminating, have good norms and be expertly tailored to participants (Kline, 2015). Additionally, a good test of witness identification accuracy should measure an individual's ability to identify the perpetrator when they are present in lineup conditions and reject a lineup when the perpetrator is not present (Megheya & Burton, 2007).

Reliability

A psychometric test is said to be reliable if all items measure the same construct (known as internal reliability). A psychometric test is also said to be reliable if the same sample obtain a consistent score when the test is repeated (known as test-retest reliability; Kline, 2015)

Internal Reliability. Internal reliability of the BFRT has been investigated using a 'split-half' model (Murray et al., 2021; Rossion & Michel, 2018). This is when a test is split into two parts and the correlation between the parts is considered (Kline, 2015). A strong corelation between the two parts of the test is indicated by a correlation coefficient of above .5 (Heale, 2015). A moderate correlation is indicated by a correlation co-efficient of .3 to .5 and a correlation below .3 is considered to be weak (Heale, 2015).

Using the split half approach, studies have found that the BFRTc has moderate internal reliability for accuracy (r =.606, Rossion & Michel, 2018) and the BFRTr has good internal reliability for accuracy (r =.735, Murray et al., 2021). Studies have also considered the inter-item correlation for the mean response time of one half of the BFRTc and BFRTr, compared to the mean response time of the other half of the BFRTc and BFRTr. Results indicated that both the BFRTc (r = .883) and BFRTr (r = .963) have good reliability for trial completion time (Murray et al., 2021; Rossion & Michel, 2018). That is, the time taken to completed one half of the test (BFRTc and BFRTr) was highly correlated to the time taken to complete the other half of the test.

Other experiments have used Cronbach's alpha to investigate internal reliability (Albonico, et al., 2017; Christensen et al., 2002; Levin et al., 1991). Cronbach's alpha provides a value of the average inter-item correlation, whereby a Cronbach's alpha of .70 or above is considered to be within the acceptable range (Kline, 2015). Using Cronbach's alpha, experiments have identified good internal reliability of the long form version of the BFRT (.72), however the internal reliability of the short form BFRT (.53) falls outside of an acceptable range as it is below .70 (Christensen et al., 2002; Levin et al., 1991). In an Italian sample, the authors calculated Cronbach's alpha and reported the internal reliability of the long form BFRT as .608, which was considered to be poor to average (Albonico et al., 2017).

Test-retest Reliability. It is important to consider test-retest reliability, as a measure that fails to yield consistent scores over time, when change in scores is not expected, is problematic (Kline, 2015). Measured using correlation analysis, it is recommended that test-retest reliability is at least .70 to be acceptable, otherwise the standard error of a test is too large for the test to be interpreted (Guildford, 1956).

Research has identified that the test-retest reliability of the long form BFRT is within an acceptable range (r =.71); however the short form reliability is outside of the acceptable range (r =.60) suggesting there may be problems with interpretability of the short form BFRT (Christensen et al., 2002; Levin et al., 1991). No data were available for the test-retest reliability of the BFRTc and the BFRTr, and it is recommended that studies collect this data so that the test-retest reliability of these tests can be considered.

Validity

The term 'validity' refers to whether a test measures what it set out to measure (Kline, 2015). There are multiple forms of validity to consider.

Concurrent Validity. The assessment of concurrent validity is concerned with the correlation of scores on a test with other tests that measure the same construct (Kline, 2015). One way to examine concurrent validity is to consider the correlation between versions of the BFRT. Benton et al. (1983) reported that the short form and long form BFRT are highly correlated for non-clinical samples (r = .88) and in clinical samples with brain disease (r = .92). The correlation between the short form and long form BFRT was supported by other researchers, who reported correlations of .88 and .84 (Albonico et al., 2017; Ferracuti, 1992). This suggests that both the short form and long form versions of the BFRT measure the same construct, presumably unfamiliar facial recognition.

Concurrent validity of the BFRTc, BFRTr and BFRT has also been the subject of research. Murray et al. (2021) reported a positive correlation between BFRTc and BFRTr that

was considered to be strong (*r* =.64). This suggests that BFRTc and BFRTr measure the same construct. Additionally, it is possible to examine concurrent validity of the BFRT by considering correlation with other tests of facial recognition such as the Cambridge Face Matching Test (CMFT; Duchaine & Nakayama, 2006). Research by Murray et al. (2021) identified the presence of a moderate correlation between the BFRTc and the CFMT (*r* =.432) and a stronger correlation between the BFRTr and the CFMT (*r* = .510). This suggests that both BFRTc and BFRTr measure face processing mechanisms, which are also measured by the CFMT, and provides further evidence of concurrent validity of the BFRT. Therefore, it appears that the BFRT, BFRTc and BFRTr display concurrent validity and measure the construct of unfamiliar facial recognition.

Predictive Validity. This is the degree to which a test can accurately predict the scores on another variable, such as how the results on an IQ test predict subsequent academic performance (Kline, 2015).

Studies have tested the predictive validity of the long form BFRT on lineup identification accuracy (Geiselman et al., 2001; Hosch, 1994; Searcy et al., 1999; Searcy et al., 2001). Hosch (1994) concluded that performance on the long form BFRT was significantly correlated to lineup identification accuracy (Experiment 1, *r* = .54; Experiment 2, *r* = .39; Experiment 3, *r* =.41; Experiment 5, *r* =.51). However there was not a significant correlation between performance on the BFRT and lineup identification accuracy in Experiment 4 (Hosch, 1994) and it is not clear what may have influenced these findings due to the limited reporting of methodological characteristics of the study. This suggests that there may be other factors that influence the predictive validity of the BFRT on lineup identification accuracy.

Furthermore, studies have found that age and lineup type impact the correlation of BFRT scores and lineup identification accuracy. In Searcy et al. (1999), senior participants

(aged 60 to 80 years) with a lower score on the BFRT had a lower accuracy (M=.18, SD=.39) on the lineup identification task than participants with a normal score on the BFRT (M=.51, SD=.57). Searcy et al. (2001) reported that it was only in target present conditions completed by 'young adult' participants (aged 18 to 30 years) in which the correlation between BFRT scores and lineup identification accuracy was statistically significant (r = .55). That is, the scores obtained by young adults on the BFRT were indicative of the accuracy of lineup identification performance when the guilty suspect was present in the lineup. However, scores obtained by older adults (aged 62 to 79 years) on the BFRT were not indicative of the accuracy of lineup performance when the guilty suspect was present in the lineup. And in target absent lineups, there was no significant correlation of BFRT scores and lineup identification accuracy for both young and older adults (Searcy et al., 2001). This suggests that age and target presence or absence impacts the relationship between BFRT scores and lineup identification accuracy. That is, scores on the BFRT are not predictive of lineup identification accuracy in older adults and target absent lineups.

Content Validity. This form of validity considers whether a test measures all aspects of a construct (Kline, 2015). Murray et al. (2021) investigated content validity by administering an inverted version of the test, whereby all images were inverted 180 degrees. This manipulation resulted in a significant difference in accuracy scores on the upright BFRTr (M= 78.83%) compared to the inverted BFRTr (M=56.66%). The authors concluded that the presence of an inversion effect supports that the BFRT measures face processing abilities. That is, because participants are impaired in their ability to complete the BFRTr when faces are inverted, this suggests that the BFRTr involves facial processing rather than image processing cognitive mechanisms, and therefore demonstrates content validity. There were no other data available for content validity of versions of BFRT.

Normative Data

Normative data, or 'norms' are sets of scores on a test from clearly defined samples, and they allow the test user to interpret scores meaningfully in relation to specific groups (Kline, 2015). Adequate norms require sufficient sampling from a large data set (Kline, 2015). The BFRT manual for the short form and long form test reports normative data for participant ages from 16 to 74 years. This normative information has been derived from data from 286 participants (including a clinical sample and non-clinical controls). Furthermore, the manual reports data from 260 children aged 6 to 14 years, which indicates that performance consistently increased with age (Benton et al., 1994). Benton and colleagues collected data from 72 elderly participants aged 75 to 84, however this data showed defective performance and was not included in the BFRT standardised norms (Benton et al., 1994). Furthermore, Albonico et al. (2017) published normative data from an Italian sample of 272 non clinical participants and 32 clinical participants (aged 19 to 31 years) with a known impairment in facial recognition. For the BFRTc, normative data are provided for 307 participants, including 202 female and 105 male aged 18 to 39 (Rossion & Michel, 2018). Finally, Murray et al. (2021) present normative data for both the BFRTc and BFRTr accuracy scores and responses times using data from 32 participants with developmental prosopagnosia. Murray et al. (2021) found that seventeen participant with known impairments in facial recognition performed within the average range. And in those that did show impairment in facial recognition ability, this was indicated by a longer task completion time Murray et al. (2021). This suggests that in addition to accuracy, response time was a valuable indicator of face recognition ability.

Sensitivity

Sensitivity is a test's ability to identify positive result, i.e., that BFRT scores indicate that an individual has face processing deficits when they do indeed have face processing

deficits. Table 8 displays the normative standards which are used to interpret participant scores on the BFRT and identify possible deficits in unfamiliar facial recognition. A corrected score of 39-40 (8^{th} to 11^{th} percentile) indicates a borderline impairment in facial recognition ability. And a corrected score of 38 or below (1^{st} to 6^{th} percentile) indicates an impairment in facial recognition ability (Benton et al., 1994). A corrected score of 41 to 49 (16^{th} to 85^{th} percentile) indicates average face recognition ability (Benton et al., 1994). A corrected score of 50 to 54 (88^{th} to 100^{th} percentile) indicates superior facial recognition ability (Benton et al., 1994).

When the short form BFRT is used, the test administrator is required to convert short form scores to long form scores using a conversion table provided (See Table 6), this allows interpretation of performance in line with long form cut offs described above (Benton et al., 1994). Furthermore, the BFRTc and BFRTr use the long form administration (Murray et al., 2021; Rossion & Michel, 2018) and so cut-off scores for the long form BFRT are applied to interpret results. Average performance by non-clinical populations on the BFRT is 45.3 (83.9%), and 44.81 (83%) on the BFRTc (Benton & Van Allen, 1968 ; Rossion & Michel, 2018). Murray et al. (2021) also measured completion times and reported that typical participants complete the BFRTr in four to six minutes and atypical participants (i.e., those with an impairment in facial recognition ability) complete the BFRTr in eight minutes.

However, it is possible for patients with known impairments in facial recognition such as prosopagnosia to score within a 'normal' range (i.e., that of a normative population). And some patients have achieved this by ignoring the identity of test faces and focussing on facial features such as eyebrows (Duchaine & Weidenfeld, 2003; Duchaine & Nakayama, 2004; Newcome, 1979; Nunn et al., 2001). On the other hand, clinical populations that have scored within a normal range were observed to take longer completing the test (Duchaine, 2000; Newcome, 1979; Nunn et al., 2001). This finding has been explained by Duchaine and

Weidenfeld (2003), who argued that the feature matching strategy employed by clinical populations, such as using eyebrows to inform facial recognition, takes more time than using a holistic processing method. It was therefore recommended that 'time-norms' are included within test-administration to improve the sensitivity of the test, as participants who take longer to complete the task may be identified as using a feature matching process to aid test performance (Duchaine & Weidenfeld, 2003).

Furthermore, Murray et al. (2021) found that the BFRTc incorrectly classified 71.88% of participants with prosopagnosia as performing within the average range, indicating poor sensitivity of the BFRTc. The BFRTr also identified 53.12% of participants with prosopagnosia as performing within the average range, suggesting that the BFRTr sensitivity was at chance level. However, the ability of the BFRTc and BFRTr to discriminate between control participants and clinical participants with prosopagnosia was also calculated using Dprime (*d'),* a bias-free measure of sensitivity whereby a score of 5 suggests perfect discriminability and 0 indicates chance discriminability (Murray et al., 2021). The results indicated that the discriminability of the BFRTr was superior (*d'* = 1.03) compared to the BFRTc (*d'* = 0.60).

Additionally, Afro-Caribbean adults obtained a slightly lower mean score (44.7) compared to the standardised Caucasian non-clinical controls (Roberts & De Hamsher, 1984). This appears to be evidence of a cross-race effect, in which there is more accurate face recognition for same race faces than for cross race faces (See Young et al., 2012 for a review). And it is likely that there is an impact of the cross-race effect on the sensitivity of all versions of the BFRT, as in each version all stimuli faces are Caucasian.

Overall, it appears that the sensitivity of the BFRT, BFRTc and BFRTr is poor, and it is important to consider time taken to complete the test. Moreover, the findings on sensitivity

of the BFRT appear to be limited to Caucasian populations, and further research is required to investigate the sensitivity of the BFRT in non-Caucasian populations.

Conclusion

This chapter argued that a test of face recognition ability may be useful to consider lineup identification outcomes, such as the accuracy of a lineup identification. In the present chapter, four versions of the BFRT were considered (Benton & Van Allen, 1968, 1994; Levin et al, 1975; Murray et al., 2021; Rossion, & Michel, 2018). There was evidence of internal reliability on the long form BFRT, BFRTc and BFRTr, but not on the short form BFRT. All versions of the BFRT demonstrated concurrent validity as they were highly correlated with each other, and other tests of facial recognition, suggesting the construct of unfamiliar facial recognition was measured. Furthermore, participants were impaired in their ability to complete the BFRTr when faces are inverted, suggesting that the BFRTr involves facial processing rather than image processing cognitive mechanisms (Murray et al., 2021). However, the literature suggests that the sensitivity of the BFRT, BFRTc and BFRTr is poor. And there appears to be a cross race effect of the BFRT (Roberts & De Hamsher, 1984). It is recommended that further research develops culturally sensitive versions of the BFRT that are appropriate for use in non-Caucasian populations.

Finally, it was identified that the long form BFRT appears to demonstrate predictive validity for eyewitness lineup identification accuracy (Geiselman et al., 2001; Hosch, 1994; Searcy et al., 1999, 2001). Notably, research has found an impact of age and lineup type (target present or absent) on the strength of the correlation between scores on the BFRT and lineup identification accuracy. Therefore, it is recommended that further research is conducted investigate the predictive validity of the short form BFRT, BFRTc and BFRTr on eyewitness lineup identification accuracy.

CHAPTER 5: DISCUSSION

Thesis Aims

The first aim of this thesis was to investigate optimal lineup construction methods that improve ability to discriminate between innocent and guilty suspects, specifically considering the impact of suspect filler similarity and distinctive facial features. In chapter 1, an overview of witness lineup identification through the lens of signal detection theory was provided and highlighted the need for lineup construction methods that enhance discriminability. Then, in chapter 2, a systematic literature review of the suspect-filler similarity research was conducted. In which, it was considered how suspect-filler similarity impacts identification responses (HR, FAR and discriminability) and the impact of methodological characteristics on experiment outcomes. Next, in chapter 3, lineup construction methods for suspects with distinctive facial features were investigated and this was likened to suspect-filler similarity dynamics as three lineup construction methods for distinctive suspects (high similarity replication, low similarity replication, do-nothing) were compared.

The second aim of this thesis was to test the feature matching model (Colloff et al., 2021) and the diagnostic feature detection theory (Wixted & Mickes, 2014) accounts of witness lineup identification decision making. In both chapters 2 and 3, the feature matching model (Colloff et al., 2021) and diagnostic feature detection theory (Wixted & Mickes, 2014) were used to make predictions about the impact of lineup construction methods on resulting suspect-filler similarity and witness ability to discriminate between innocent and guilty suspects.

The third aim of this thesis was to evaluate the use of the Benton Facial Recognition Test (BFRT) and its applicability to considering individual differences in witness identification performance and this is detailed in chapter 4.

Finally, the current chapter will explore and discuss the findings of the thesis and outline implications for theory and practice.

Theoretical Predictions

The feature matching model (Colloff et al., 2021) and diagnostic feature detection theory (Wixted & Mickes, 2014) were used to make predictions about optimal filler similarity in fair lineups (where all fillers match the witness's description of the perpetrator) and unfair lineups (where the suspect stands out).

Feature Matching Model (Colloff et al., 2021)

In fair *target present conditions*, higher similarity fillers will share many features that match the witness's memory of the perpetrator, reducing the number of unique features on the perpetrator in the lineup that match the witness's memory of the perpetrator. That is, higher similarity fillers compete with the witness's memory of the perpetrator, resulting in a decrease in the hit rate. Lower similarity fillers will share fewer features that match the witness's memory of the perpetrator, increasing the number of unique features on the perpetrator that match the witness's memory of the perpetrator. That is, lower similarity fillers compete with the witness's memory of the perpetrator to a lesser extent, resulting in an increase in the hit rate. In fair *target absent conditions,* where filler similarity to the innocent suspect has been manipulated (i.e., suspect matched lineups), varying filler similarity should not change the number of features on the face of the innocent suspect that match the witness's memory of the perpetrator. Therefore, the false alarm rate should remain unchanged across lower and higher filler similarity conditions. In unfair *target absent conditions,* varying filler similarity will cause the innocent suspect to stand out in memory to a greater extent in lower similarity conditions, because the innocent suspect shares more features with the perpetrator in memory than do the other fillers, and therefore the false alarm rate will increase.

Diagnostic Feature Detection Theory (Wixted & Mickes, 2014)

In fair *high similarity lineups* the availability of potential diagnostic features is reduced due to the resemblance between the fillers and perpetrator, reducing discriminability. In fair *low similarity lineups* all lineup members will share some non-diagnostic features that have been used to match the fillers to the perpetrator or innocent suspect, however there are still diagnostic features available to make an identification decision, increasing discriminability. In *unfair lineups,* when the innocent suspect resembles the perpetrator more than other fillers, it is not clear that certain features on the face of the innocent suspect are non-diagnostic of guilt. As such, the witness is more likely to focus on non-diagnostic features to make an identification decision, decreasing discriminability.

Impact of Methodological Characteristics

Our results highlight the importance of methodological characteristics in lineup construction methods. In chapter 2, a systematic literature review was conducted, where identification response outcomes for twenty nine experiments manipulating filler similarity were predicted. It was found that there were no standardised procedures for constructing lineups within the filler similarity literature. This created methodological inconsistencies in filler selection (i.e., suspect or perpetrator matched, match to appearance or description or both), and innocent suspect selection (i.e., highly similar, description matched, moderately similar or dissimilar). According to signal-detection based models, these methodological differences influence the overall similarity between the fillers and the perpetrator; fillers and the innocent suspect; fillers and the perpetrator; and the innocent suspect and the perpetrator. Consequently, there was not a consistent pattern across experiments in the hit rate, false alarm rate and the discriminability measure across lineups of varying similarity. Put another way, low similarity lineups did not always optimise lineup performance.

Although low similarity lineups did not always optimise lineup performance, this was expected due to differences in methodological decisions made across experiments. The assumptions of the feature matching model and diagnostic feature detection theory (Colloff et al., 2021; Wixted & Mickes, 2014) were used to make predictions, explain existing results, and provide recommendations for achieving optimal lineup construction methods to enhance discriminability. According to a signal detection interpretation of a fair lineup, there are two memory distributions within witness memory, one for the perpetrator and the other for fillers and the innocent suspect (Colloff et al., 2021; Wixted & Mickes, 2014). Any methodological decision that results in the innocent suspect standing out as more similar to the witness's memory of the perpetrator than the other lineup members, theoretically results in three memory strength distributions (see figure 4). That is, there will be an overlap of the innocent suspect distribution and the perpetrator distribution to a greater extent than the filler distribution overlaps with the perpetrator distribution. This will subsequently impact ability to discriminate between innocent and guilty suspects and the superiority of the low similarity condition in target absent conditions. In low similarity conditions, fillers are dissimilar to the suspect, but within the constraints of the witness description. So, the innocent suspect stands out due to being more similar to the guilty suspect than the fillers. Therefore, the lineup becomes unfair and there will be more false alarms to the innocent suspect.

Filler and Innocent Suspect Selection

A source of methodological variance in experiments is how fillers are selected. That is, whether fillers are matched to the perpetrator in both target present and target absent conditions (i.e., perpetrator matched) or matched to the perpetrator in target present conditions and matched to the innocent suspect in target absent conditions (i.e., suspect matched). When fillers are suspect matched, varying filler similarity should not change the number of features on the face of the innocent suspect that match the witness's memory of

the perpetrator. Therefore, the false alarm rate should remain unchanged across lower and higher filler similarity conditions (Colloff et al., 2021; Wixted & Mickes, 2014). When fillers are perpetrator matched, varying filler similarity will cause the innocent suspect to stand out in memory in lower similarity conditions, because the innocent suspect shares more features with the perpetrator in memory than do the other fillers, and therefore the false alarm rate will increase. Moreover, fewer shared features can be discounted when lower similarity compared to higher similarity fillers are used, reducing discriminability in low similarity compared to high similarity lineups (Colloff et al., 2021; Wixted & Mickes, 2014).

Another source of methodological variation is how the innocent suspect is selected. From a signal-detection framework (Colloff et al., 2021; Wixted & Mickes, 2014), using an innocent suspect who is more similar to the witness's memory of the perpetrator than the fillers, on average theoretically results in three memory distributions. One for the perpetrator, one for the innocent suspect who shares a higher proportion of the perpetrators features than the fillers, and one distribution for the fillers who have less of the perpetrator's unique features. That is, presenting an innocent suspect who is highly similar to the perpetrator along with moderate similarity or low similarity fillers, results in an unfair lineup.

In chapter 2, filler and innocent suspect selection were considered when theoretical models (Colloff et al., 2021; Wixted & Mickes, 2014) were applied to predict patterns of the HR, FAR and discriminability with reasonable success. It was argued that a low similarity lineup advantage exists, when lineup construction methods result in the innocent suspect being no more similar to the perpetrator than the other lineup members. And this may be achieved by avoiding the use of high similarity innocent suspects and using suspect matched filler selection methods (Clark & Tunnicliff, 2001; Colloff et al., 2021; Oriet & Fitzgerald, 2018). These findings challenge previous recommendations that fillers should be moderately to highly similar to the suspect in order to present the innocent suspect from misidentification

(Fitzgerald et al., 2013, 2015). It is argued that constructing lineups to include lower similarity fillers (including those with distinctive facial features) does not put the innocent suspect at increased risk of misidentification but does increase the number of correct identifications of the guilty suspect (Colloff et al., 2021). Henceforth, this method of lineup construction is in keeping with guidance to increase discriminability (NRC, 2014) while ensuring that the lineup remains fair (i.e., the suspect does not stand out).

Distinctive Suspects

In chapter 3, lineup construction methods for suspects with distinctive facial features was considered and likened this to suspect-filler similarity dynamics. Theoretical predictions (Colloff et al., 2021; Wixted & Mickes, 2014) were tested by comparing lineup construction methods (high similarity replication, low similarity replication and do-nothing) for distinctive suspects. Suspect matched fillers were used and an innocent suspect who was of median similarity to the perpetrator, compared to the other fillers. Unlike previous studies (Colloff et al., 2017, 2017), our innocent suspect had a different distinctive feature of a star tattoo, compared to the perpetrator's tribal tattoo. That is, our innocent suspect was of median similarity to the perpetrator, but the distinctive feature fit the general description of the perpetrator ('tattoo on the face').

Indeed, the pattern of results supported theoretical predictions (Colloff et al., 2021; Wixted & Mickes, 2014). In low similarity replication compared to high similarity replication lineups, there was an increase in the HR, no difference in FAR and an increase in discriminability. That is, low-similarity fair lineups, in which fillers are matched to the suspect, did not put the innocent suspect at increased risk of being falsely identified (Colloff et al., 2021; Oriet & Fitzgerald, 2018).

It was also predicted that ability to discriminate innocent from guilty suspects would be lower in the high similarity replication lineups than do-nothing replication lineups. This

was a weak prediction, because it was based on the (untested) premise that participants would make proper use of the of the diagnostic information provided by the tattoos in the do-nothing condition (Wixted & Mickes, 2014). As the innocent suspect's tattoo was not an exact match to the perpetrator's (e.g., the innocent suspect has a star tattoo, while the perpetrator has a tribal tattoo), it was assumed that participants would scrutinise the shape and design of the tattoo and only make an identification of the suspect if the tattoo exactly matches memory.

Contrary to predictions, ability to discriminate innocent from guilty suspects was not better in do-nothing lineups compared to high similarity lineups. That is, when presented with a lineup in which the suspect stood out as being the only person with a distinctive facial tattoo, participants were not better able to discriminate between innocent and guilty suspects than when they were faced with a lineup in which all lineup members had the same distinctive facial feature. It was argued that separate processes were probably underlying the seemingly equivalent performance. In the high similarity replication condition, it was likely that participants were (as predicted) theoretically discounting of a useful feature (tattoo) because it was shared by all lineup members. Whereas, in do-nothing conditions, it appears that participants were glossing over the distinctive feature characteristics and using the mere existence of a facial tattoo as evidence of a memory match, regardless of the guilt or innocence of the suspect. As a result, both processes result in an overall decrease in the ability to discriminate between innocent and guilty suspects compared to the low similarity replication condition. It appears that scrutinising the facial tattoo in the low similarity replication conditions results in improved ability to discriminate between innocent and guilty suspects, compared to do-nothing and high similarity replication conditions.

The results of chapter 3 have implications for lineup construction for distinctive suspects in practice. This chapter presented the first empirical test of high and low similarity

replication methods for distinctive suspects. It was argued that, when constructing lineups for distinctive suspects, replicating a similar but non identical distinctive feature across other lineup members will result in an improved ability to discriminate between innocent and guilty suspects**.** However, the constraints on generality should be noted (Simons et al., 2017), as the present results are limited to conditions in which the witness has provided a general description of the distinctive facial feature (i.e., 'tattoo on the face'). It is advised that further research should seek to replicate our findings before recommendations are made to the criminal justice system.

Individual Differences

It was argued that the use of facial recognition tests within eyewitness research may be beneficial to evaluate if individual differences in facial recognition ability exist and are associated with lineup performance. In chapter 4, four versions of the BFRT (short form, long form, BFRTc, BFRTr) were considered as a measure of facial recognition ability and an indicator of lineup identification performance. It was found that scores on a short form BFRT did not appear to demonstrate acceptable reliability. Whereas all other versions of the BFRT did demonstrate acceptable reliability. Additionally, all versions of the BFRT were highly correlated with each other, and other tests of facial recognition, suggesting the construct of unfamiliar facial recognition was measured. However, the sensitivity of the BFRT, BFRTc and BFRTr is poor (Duchaine & Weidenfeld, 2003; Duchaine & Nakayama, 2004; Murray et al., 2021; Newcome, 1979; Nunn et al., 2001). That is, the ability of the BFRT, BFRTc and BFRTr to identify an individual as having a deficit in facial recognition ability, when they do indeed have a deficit in facial recognition ability is poor. Moreover, there appears to be a cross race effect of the BFRT long form, as non-Caucasian test completers score lower than Caucasian test completers (Roberts & De Hamsher, 1984). It is recommended further

research is required to develop culturally sensitive versions of the BFRT that are appropriate for use in non-Caucasian populations.

Finally, it was identified that the long form BFRT appears to demonstrate predictive validity for eyewitness lineup identification accuracy (Geiselman et al., 2001; Hosch, 1994; Searcy et al., 1999, 2001). Notably, research has found an impact of age and lineup type (target present or absent) on the strength of the correlation between scores on the BFRT and lineup identification accuracy. Therefore, it is recommended that further research is conducted to investigate the predictive validity of the short form BFRT, BFRTc and BFRTr on eyewitness lineup identification accuracy. Moreover, it is recommended that future research could also investigate if there is a relationship between suspect-filler similarity (i.e., low, and high), facial recognition ability (i.e., poor to super recogniser) and lineup identification accuracy. This is an area that has not previously been considered within the literature and could add to our understanding of the interaction of filler similarity and lineup identification accuracy when individual differences such as facial recognition ability are considered.

Conclusion

The present thesis sought to disentangle findings in the literature and provide new direction within the field to establish optimal lineup construction methods. It is recommended that psychometric tools are used to allow for further exploration of lineup construction methods that enhance discriminability when individual differences are also considered. It is also recommended that more research is conducted to establish the usefulness of the BFRT as an indicator of individual differences in lineup identification (see chapter 4). Moreover, it is argued that it has not been possible to identify optimal lineup construction methods that consistently increase discriminability across lineup conditions to date because of methodological variances in existing literature (see chapter 2) and within policing practice

worldwide (Fitzgerald et al., 2021). Consistent with theoretical predictions, the findings of chapters 2 and 3 demonstrate that low similarity lineups allow the witness to focus on the perpetrators' unique features that are diagnostic of guilt and make an accurate identification decision (Colloff et al., 2021: Wixted & Mickes, 2014). However, the low similarity lineup advantage holds only when lineup construction methods result in lineups in which the suspect does not stand out (i.e., suspect matched and moderate similarity innocent suspect). Consistent with Quigley-McBride and Wells (2021), it is recommended that researchers thoroughly report lineup construction methods and make experimental materials and data publicly available. When further studies have been conducted, it is recommended that a meta-analysis is completed to further test optimal lineup conditions as a function of experimental methodology.

References

Experiments included in the systematic literature review are marked with two asterisks **

Albonico, A., Malaspina, M., & Daini, R. (2017). Italian normative data and validation of two neuropsychological tests of face recognition: Benton Facial Recognition Test and Cambridge Face Memory Test. Neurological Sciences, 38(9), 1637-1643. https://doi-org.ezproxye.bham.ac.uk/10.1007/s10072-017-3030-6

Andersen, S. M., Carlson, C. A., Carlson, M. A., & Gronlund, S. D. (2014). Individual differences predict eyewitness identification performance. Personality and Individual Differences, 60, 36-40. https://doi.org/10.1016/j.paid.2013.12.011

Badham, S. P., Wade, K. A., Watts, H. J., Woods, N. G., & Maylor, E. A. (2013). Replicating distinctive facial features in lineups: Identification performance in young versus older adults. Psychonomic bulletin & review, 20(2), 289-295. https://doi.org/10.3758/s13423-012-0339-2

Baldassari, M. J., Kantner, J., & Lindsay, D. S. (2019). The importance of decision bias for predicting eyewitness lineup choices: toward a Lineup Skills Test. Cognitive research: principles and implications, 4(1), 1-13.https://doi.org/10.1186/s41235-018-0150-3

Barton, J. J., & Corrow, S. L. (2016). The problem of being bad at faces. Neuropsychologia, 89, 119-124. https://doi.org/10.1016/j.neuropsychologia.2016.06.008

Bate, S., Frowd, C., Bennetts, R., Hasshim, N., Murray, E., Bobak, A. K., ... & Richards, S. (2018). Applied screening tests for the detection of superior face recognition. Cognitive research: principles and implications, 3(1), 1-19. https://doi.org/10.1186/s41235-018-0116-5

Bate, S., Frowd, C., Bennetts, R., Hasshim, N., Portch, E., Murray, E., & Dudfield, G. (2019). The consistency of superior face recognition skills in police officers. Applied Cognitive Psychology, 33(5), 828-842. https://doi.org/10.1002/acp.3525

Benton, A. L., & Van Allen, M. W. (1968). Impairment in facial recognition in patients with cerebral disease. Cortex, 4(4), 344-IN1. https://doi.org/10.1016/S0010-9452(68)80018-8

Benton, A. L., & Van Allen, M. W. (1972). Prosopagnosia and facial discrimination. Journal of the neurological sciences, 15(2), 167-172. https://doi.org/10.1016/0022-510X(72)90004-4

Benton, A. L., Sivan, A. B., deS, K., Varney, N. R., & Spreen, O. (1983). Facial Recognition: Stimulus and Multiple Choice Pictures: Contributions to Neuropsychological Assessment. Oxford University Press, Incorporated.

Benton, A. L., Sivan, A. B., Hamsher, K. D., Varney, N. R., & Spreen, O. (1994). Contributions to neuropsychological assessment: A clinical manual. Oxford University Press, USA.

Bergold, A. N., & Heaton, P. (2018). Does Filler Database Size Influence Identification Accuracy? Law & Human Behavior, 42(3), 227-243. https:// doi.org/10.1037/lhb0000289 **

Bindemann, M., Brown, C., Koyas, T., & Russ, A. J. (2012). Individual differences in face identification postdict eyewitness accuracy. Journal of Applied Research in Memory and Cognition, 1, 96–103. https://doi.org/10.1016/j.jarmac.2012.02.001.

Booth, S. (2019). The Dud Effect: The Effect of Dissimilar Fillers in Eyewitness Lineups. (M.A.). City University of New York John Jay College of Criminal Justice, Ann Arbor.

Bora, E., Eryavuz, A., Kayahan, B., Sungu, G., & Veznedaroglu, B. (2006). Social functioning, theory of mind and neurocognition in outpatients with schizophrenia; mental state decoding may be a better predictor of social functioning than mental state

reasoning. Psychiatry Research, 145(23), 95103. https://doi.org/10.1016/j.psychres.2005.11.003

Bowles, D. C., McKone, E., Dawel, A., Duchaine, B., Palermo, R., Schmalzl, L., ... & Yovel, G. (2009). Diagnosing prosopagnosia: Effects of ageing, sex, and participant–stimulus ethnic match on the Cambridge Face Memory Test and Cambridge Face Perception Test. Cognitive neuropsychology, 26(5), 423-455. https://doi.org/10.1080/02643290903343149

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

Broks, P., Young, A. W., Maratos, E. J., Coffey, P. J., Calder, A. J., Isaac, C. L., . . . Hadley, D. (1998). Face processing impairments after encephalitis: Amygdala damage and recognition of fear. Neuropsychologia, 36(1), 59-70. doi:http://dx.doi.org/10.1016/S0028-3932%2897%2900105-X

Bruyer, R., & Schweich, M. (1991). A clinical test battery of face processing. International Journal of Neuroscience, 61(1-2), 19–30. https://doi.org/10.3109/00207459108986268

Burton, A. M., White, D., & McNeill, A. (2010). The Glasgow face matching test. Behavior research methods, 42(1), 286-291.

Bushman, B.J., & Wang, M. C. (2009). Vote-Counting Procedures in Meta-Analysis. In Cooper, H., Hedges, L.V., & Valentine, J.C. (Eds), *The Handbook of Research Synthesis and Meta-Analysis* (pp. 207-220). New York: Russell Sage Foundation.

Busigny, T., Van Belle, G., Jemel, B., Hosein, A., Joubert, S., & Rossion, B. (2014). Face-specific impairment in holistic perception following focal lesion of the right anterior temporal lobe. Neuropsychologia, 56, 312-333. https://doi.org/10.1016/j.neuropsychologia.2014.01.018

Calder, A. J., Keane, J., Manes, F., Antoun, N., & Young, A. W. (2000). Impaired recognition and experience of disgust following brain injury. Nature Neuroscience, 3(11), 1077-1078. doi:http://dx.doi.org/10.1038/80586

Carlson, C. A., & Carlson, M. A. (2014). An evaluation of lineup presentation, weapon presence, and a distinctive feature using ROC analysis. Journal of Applied Research in Memory and Cognition, 3(2), 45-53. https://doi.org/10.1016/j.jarmac.2014.03.004

Carlson, C. A., Gronlund, S. D., & Clark, S. E. (2008). Lineup Composition, Suspect Position, and the Sequential Lineup Advantage. Journal of Experimental Psychology: Applied, 14(2), 118-128. https://doi.org/10.1037/1076-898X.14.2.118**

Chance, J. E., & Goldstein, A. G. (1996). The other-race effect and eyewitness identification. In S. L. Sporer, R. S. Malpass, & G. Koehnken (Eds.), Psychological issues in eyewitness identification (p. 153–176). Lawrence Erlbaum Associates, Inc.

Charman, S. D., Wells, G. L., & Joy, S. W. (2011). The dud effect: Adding highly dissimilar fillers increases confidence in lineup identifications. Law and human behavior, 35(6), 479-500. https://doi.org/10.1007/s10979-010-9261-1

Cheung, M. W. L., & Vijayakumar, R. (2016). A guide to conducting a meta-analysis. Neuropsychology review, 26(2), 121-128. https://doi.org/10.1007/s11065-016-9319-z

Christensen, K. J., Riley, B. E., Heffernan, K. A., Love, S. B., & McLaughlin Sta. Maria, M. E. (2002). Facial recognition test in the elderly: Norms, reliability, and premorbid estimation. The Clinical Neuropsychologist, 16(1), 5156. https://doi.org/10.1076/clin.16.1.51.8332

Clark, S. E. (2005). A Re-examination of the Effects of Biased Lineup Instructions in Eyewitness Identification. Law and Human Behavior, 29(4), 395–424. https://doi.org/10.1007/s10979-005-5690-7

Clark, S. E. (2012). Costs and benefits of eyewitness identification reform: Psychological science and public pol- icy. Perspectives on Psychological Science, 7, 238–259. doi:10.1177/1745691612439584

Clark, S. E., & Tunnicliff, J. L. (2001). Selecting Lineup Foils in Eyewitness Identification Experiments: Experimental Control and Real-World Simulation. Law & Human Behavior, 25(3), 199-216. https://doi.org/10.1023/A:1010753809988

Clark, S.E., Moreland, M.B. & Gronlund, S.D.(2014). Evolution of the empirical and theoretical foundations of eyewitness identification reform. Psychonomic Buletin and Review, 21, 251–267. https://doi.org/10.3758/s13423-013-0516-y

Clarke, S., Lindemann, A., Maeder, P., Borruat, F. X., & Assal, G. (1997). Face recognition and postero-inferior hemispheric lesions. Neuropsychologia, 35(12), 1555-1563. doi:10.1016/S0028-3932(97)00083-3

Cohen, A. L., Starns, J. J., Rotello, C. M., & Cataldo, A. M. (2020). Estimating the proportion of guilty suspects and posterior probability of guilt in lineups using signal-detection models. Cognitive research: principles and implications, 5(1). https://doi.org/10.1186/s41235-020-00219-4

Colloff., M. F. (2016). *Eyewitness identification performance on lineups for distinctive suspects.* PhD thesis, University of Warwick. http://wrap.warwick.ac.uk/90153

Colloff, M. F., & Wixted, J. T. (2020). Why are lineups better than showups? A test of the filler siphoning and enhanced discriminability accounts. Journal of Experimental Psychology: Applied, 26(1), 124. https://psycnet.apa.org/doi/10.1037/xap0000218

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. https://doi.org/10.1177/0956797616655789 **

Colloff, M. F., Wade, K. A., Strange, D., & Wixted, J. T. (2018). Filler-siphoning theory does not predict the effect of lineup fairness on the ability to discriminate innocent from guilty suspects: Reply to Smith, Wells, Smalarz, and Lampinen (2018). Psychological Science, 29(9), 1552–1557. https://doi.org/10.1177/0956797618786459

Colloff, M. F., Wade, K. A., Wixted, J. T., & Maylor, E. A. (2017). A signal-detection analysis of eyewitness identification across the adult lifespan. Psychol Aging, 32(3), 243-258. https://doi.org/10.1037/pag0000168 **

Colloff, M. F., Wilson, B. M., Seale-Carlisle, T. M., & Wixted, J. T. (2021). Optimizing the selection of fillers in police lineups. Proceedings of the National Academy of Sciences, 118(8). https://doi.org/10.1073/pnas.2017292118**

Cutler, B. L., Penrod, S. D., & Martens, T. K. (1987). Improving the Reliability of Eyewitness Identification: Putting Context Into Context. Journal of Applied Psychology, 72(4), 629-637. https://doi.org/10.1037/0021-9010.72.4.629

Darling, S., Martin, D., Hellmann, J. H., & Memon, A. (2009). Some witnesses are better than others. Personality and Individual Differences, 47(4), 369-373. https://doi.org/10.1016/j.paid.2009.04.010

Darling, S., Valentine, T., & Memon, A. (2008). Selection of lineup foils in operational contexts. Applied Cognitive Psychology, 22(2), 159-169. https://doi.org/10.1002/acp.1366

De Renzi, E. (1986). Current issues on prosopagnosia. In Aspects of face processing (pp. 243-252). Springer, Dordrecht.

Deffenbacher, K. A., Bornstein, B. H., Penrod, S. D., & McGorty, E. K. (2004). A meta-analytic review of the effects of high stress on eyewitness memory. Law and human behavior, 28(6), 687-706. https://doi.org/10.1007/s10979-004-0565-x

Demirci, E., & Erdogan, A. (2016). Is emotion recognition the only problem in ADHD? Effects of pharmacotherapy on face and emotion recognition in children with ADHD. ADHD Attention Deficit and Hyperactivity Disorders, 8(4), 197-204. https://doi.org/10.1007/s12402-016-0201-x

Devenport, J. L., & Cutler, B. L. (2004). Impact of Defense-Only and Opposing Eyewitness Experts on Juror Judgments. Law & Human Behavior, 28(5), 569-576. https://doi.org/10.1023/B:LAHU.0000046434.39181.07

Dianiska, R. E., Manley, K. D., & Meissner, C. A. (2021). A Process Perspective: The Importance of Theory in Eyewitness Identification Research. Methods, Measures, and Theories in Eyewitness Identification Tasks, 136-168.

Duchaine, B. C. (2000). Developmental prosopagnosia with normal configural processing. Neuroreport, 11(1), 79-83.

Duchaine, B. C., & Nakayama, K. (2004). Developmental prosopagnosia and the Benton Facial Recognition Test. Neurology, 62(7), 1219–1220. https://doi.org/10.1212/01.WNL.0000118297.03161.B3.

Duchaine, B. C., & Weidenfeld, A. (2003). An evaluation of two commonly used tests of unfamiliar face recognition. Neuropsychologia, 41(6), 713-720. https://doi.org/10.1016/S0028-3932(02)00222-1

Duchaine, B., & Nakayama, K. (2006). The Cambridge Face Memory Test: Results for neurologically intact individuals and an investigation of its validity using inverted face stimuli and prosopagnosic participants. Neuropsychologia, 44(4), 576-585. https://doi.org/10.1016/j.neuropsychologia.2005.07.001

Duchaine, B., Germine, L., & Nakayama, K. (2007). Family resemblance: Ten family members with prosopagnosia and within-class object agnosia. Cognitive neuropsychology, 24(4), 419-430. https://doi.org/10.1080/02643290701380491

Eisen, M. L., Gabbert, F., Ying, R., & Williams, J. (2017). "I think he had a tattoo on his neck": How co-witness discus- sions about a perpetrator's description can affect eyewitness identification decisions. Journal of Applied Research in Memory and Cognition, 6, 274–282. https://doi.org/10.1016/j.jarmac.2017.01.009

Ferracuti F, Ferracuti S (1992) Taratura del campione italiano. In:test di riconoscimento di volti ignoti. Firenze, p 26–29

Fitzgerald, R. J., Oriet, C., & Price, H. L. (2015). Suspect filler similarity in eyewitness lineups: A literature review and a novel methodology. Law and human behavior, 39(1), 62-74. https://doi.org/10.1037/lhb0000095 **

Fitzgerald, R. J., Price, H. L., Oriet, C., & Charman, S. D. (2013). The effect of suspect-filler similarity on eyewitness iden- tification decisions: A meta-analysis. Psychology, Public Policy, and Law, 19, 151–164. https://doi.org/10.1037/a0030618

Fitzgerald, R. J., Rubínová, E., & Juncu, S. (2021). Eyewitness identification around the world. In Methods, measures, and theories in eyewitness identification tasks (pp. 294-322). Routledge.

Flowe, H. D., & Ebbesen, E. B. (2007). The effect of lineup member similarity on recognition accuracy in simultaneous and sequential lineups. Law and Human Behavior, 31(1), 33-52. https://doi.org/10.1007/s10979-006-9045-9**

Flowe, H. D., Carline, A., & Karoğlu, N. (2018). Testing the reflection assumption: A comparison of eyewitness ecology in the laboratory and criminal cases. The International Journal of Evidence & Proof, 22(3), 239–261. https://doi.org/10.1177/1365712718782996

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. https://doi.org/10.1037/lhb0000101

Garg, A. X., Hackam, D., & Tonelli, M. (2008). Systematic review and meta-analysis: when one study is just not enough. Clinical Journal of the American Society of Nephrology, 3(1), 253-260. https://doi.org/10.2215/CJN.01430307

Geiselman, R. E., Tubridy, A., Bkynjun, R., Schroppel, T., Turner, L., Yoakum, K., & Young, N. (2001). Benton Facial Recognition Test scores: Index of eyewitness accuracy. American Journal of Forensic Psychology.

Gettleman, J. N. , Grabman, J. H. , Dobolyi, D. G. & Dodson, C. S. (2021). A Decision Processes Account of the Differences in the Eyewitness Confidence-Accuracy Relationship Between Strong and Weak Face Recognizers Under Suboptimal Exposure and Delay Conditions. Journal of Experimental Psychology: Learning, Memory, and Cognition, 47 (3), 402-421. doi: 10.1037/xlm0000922.

Gonzalez, R., Davis, J., & Ellsworth, P. C. (1995). Who should stand next to the suspect? Problems in the assessment of lineup fairness. Journal of Applied Psychology, 80(4), 525. https://doi.org/10.1037/0021-9010.80.4.525

Grabman, J. H., & Dodson, C. S. (2020). Stark individual differences: Face recognition ability influences the relationship between confidence and accuracy in a recognition test of Game of Thrones actors. Journal of Applied Research in Memory and Cognition, 9(2), 254-269. https://doi.org/10.1016/j.jarmac.2020.02.007

Grabman, J. H., Dobolyi, D. G., Berelovich, N. L., & Dodson, C. S. (2019). Predicting high confidence errors in eyewitness memory: The role of face recognition ability, decision-time, and justifications. Journal of Applied Research in Memory and Cognition, 8(2), 233-243. https://doi.org/10.1016/j.jarmac.2019.02.002

Green, D. M., & Swets, J. A. (1966). Signal detection theory and psychophysics (Vol. 1). New York: Wiley.

Gronlund, S. D., Carlson, C. A., Dailey, S. B., & Goodsell, C. A. (2009). Robustness of the sequential lineup advantage. Journal of Experimental Psychology: Applied, 15, 140–152. https://doi.org/10.1037/a0015082 **

Guilford, J. P. (1956). The structure of intellect. Psychological bulletin, 53(4), 267.

Heale, R., & Twycross, A. (2015). Validity and reliability in quantitative studies. Evidence-based nursing, 18(3), 66-67.

Herlitz, A., & Lovén, J. (2013). Sex differences and the own-gender bias in face recognition: A meta-analytic review. Visual Cognition, 21(9-10), 1306-1336. https://doi.org/10.1080/13506285.2013.823140

Horry, R., & Brewer, N. (2016). How Target-Lure Similarity Shapes Confidence Judgments in Multiple-Alternative Decision Tasks. Journal of Experimental Psychology: General, 145(12), 1615-1634. https://doi.org/10.1037/xge0000227 **

Hosch, H. (1994). Individual differences in personality and eyewitness identification. In D. Ross, J. Read, & M. Toglia (Eds.), Adult Eyewitness Testimony: Current Trends and Developments (pp. 328-347). Cambridge: Cambridge University Press. doi:10.1017/CBO9780511759192.016

Innocence Project (2022). Eyewitness Identification Reform. Retrieved from: https://innocenceproject.org/eyewitness-identification-reform/

Juslin, P., Olsson, N., & Winman, A. (1996). Calibration and Diagnosticity of Confidence in Eyewitness Identification: Comments on What Can Be Inferred From the Low Confidence-Accuracy Correlation. Journal of Experimental Psychology: Learning, Memory, & Cognition, 22(5), 1304-1316. https://doi.org/10.1037/0278-7393.22.5.1304**

Key, K. N., Cash, D. K., Neuschatz, J. S., Price, J., Wetmore, S. A., & Gronlund, S. D. (2015). Age differences (or lack thereof) in discriminability for lineups and showups. Psychology, Crime & Law, 21(9), 871-889. https://doi.org/10.1080/1068316X.2015.1054387 **

Key, K. N., Wetmore, S. A., Neuschatz, J. S., Gronlund, S. D., Cash, D. K., & Lane, S. (2017). Line-up Fairness Affects Postdictor Validity and 'Don't Know' Responses. Applied Cognitive Psychology, 31(1), 59-68. https://dx.doi.org/10.1002/acp.3302 **

Kline, P. (2015). A handbook of test construction (psychology revivals): introduction to psychometric design. Routledge.

Lancelot, C., & Gilles, C. (2019). How does visual context influence recognition of facial emotion in people with traumatic brain injury? Brain Injury, 33(1), 4-11. doi:http://dx.doi.org/10.1080/02699052.2018.1531308

Levi, A. (2016). Once again: selecting foils as similar to the suspect, or matching the description of the culprit? Journal of Criminal Psychology, 6(3), 114-120. https://doi.org/10.1108/JCP-03-2016-0011

Levin, B., Llabre, M., Reisman, S., Weiner, W., Sanchez-Ramos, J., Singer, C., & Brown, M. (1991). Visuospatial impairment in Parkinson's disease. Neurology, 41(3), 365–365. https://doi.org/10.1212/WNL.41.3.365

Levin, H. S., Hamsher, K. D. S., & Benton, A. L. (1975). A short form of the test of facial recognition for clinical use. The Journal of Psychology, 91(2), 223-228. https://doi.org/10.1080/00223980.1975.9923946

Lindsay, D. S., Read, J. D., & Sharma, K. (1998). Accuracy and confidence in person identification: The relationship is strong when witnessing conditions vary widely. Psychological Science, 9, 215-218. https://doi.org/10.1111%2F1467-9280.00041

Lindsay, R. C. L., & Wells, G. L. (1980). What price justice? - Exploring the relationship of lineup fairness to identification accuracy. Law and human behavior, 4(4), 303-313. https://doi.org/10.1007/BF01040622**

Lindsay, R. C. L., Lea, J. A., Nosworthy, G. J., Fulford, J. A., Hector, J., LeVan, V., & Seabrook, C. (1991). Biased Lineups: Sequential Presentation Reduces the Problem. Journal of Applied Psychology, 76(6), 796-802. https://doi.org/10.1037/0021-9010.76.6.796

Lindsay, R. C. L., Lea, J. A., Nosworthy, G. J., Fulford, J. A., Hector, J., LeVan, V., & Seabrook, C. (1991). Biased lineups: Sequential presentation reduces the problem. Journal of Applied Psychology, 76(6), 796–802. https://doi.org/10.1037/0021-9010.76.6.796

Lindsay, R. C. L., Martin, R., & Webber, L. (1994). Default Values in Eyewitness Descriptions: A Problem for the Match-to-Description Lineup Foil Selection Strategy. Law & Human Behavior, 18(5), 527-541. https://doi.org/10.1007/BF01499172

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://doi.org/10.1037/0021-9010.70.3.556

Lindsay, R. C., Ross, D. F., Read, J. D., & Toglia, M. P. (2007). The handbook of eyewitness psychology. Volume II: Memory for People. London: LEA.

Loftus, E. F., Loftus, G. R., & Messo, J. (1987). Some facts about “weapon focus”. Law and human behavior, 11(1), 55-62. https://doi.org/10.1007/BF01044839.

Lovén, J., Herlitz, A., & Rehnman, J. (2011). Women’s own-gender bias in face recognition memory. Experimental psychology. https://doi.org/10.1027/1618-3169/a000100

Lucas, C. A., & Brewer, N. (2021). Could precise and replicable manipulations of suspect-filler similarity optimize eyewitness identification performance?. Psychology, Public Policy, and Law. https://psycnet.apa.org/doi/10.1037/law0000329**

Lucas, C. A., Brewer, N., & Palmer, M. A. (2021). Eyewitness identification: The complex issue of suspect-filler similarity. Psychology, Public Policy, and Law, 27(2), 151–169. https://doi.org/10.1037/law0000243**

Luus, C. A., & Wells, G. L. (1994). The malleability of eyewitness confidence: Co-witness and perseverance effects. Journal of Applied Psychology, 79 (5), 714-723. https://doi.org/10.1037/0021-9010.79.5.714

Luus, C. E., & Wells, G. L. (1991). Eyewitness identification and the selection of distracters for lineups. Law and Human Behavior, 15(1), 43-57. https://doi.org/10.1007/BF01044829

Macmillan, N. A., & Creelman, C. D. (2004). Detection theory: A user's guide. Psychology press.

Megreya, A. M., & Burton, A. M. (2007). Hits and false positives in face matching: A familiarity-based dissociation. Perception & Psychophysics, 69, 1175–1184. https://doi.org/10.3758/BF03193954.

Memon, A., Hope, L., & Bull, R. (2003). Exposure duration: Effects on eyewitness accuracy and confidence. British Journal of Psychology, 94(3), 339-354.https://doi.org/10.1348/000712603767876262

Mickes, L. (2015). Receiver operating characteristic analysis and confidence-accuracy characteristic analysis in investigations of system variables and estimator variables that affect eye- witness memory. Journal of Applied Research in Memory and Cognition, 4, 93–102. doi:10.1016/j.jarmac.2015.01.003

Mickes, L., Seale-Carlisle, T. M., Wetmore, S. A., Gronlund, S. D., Clark, S. E., Carlson, C. A., Goodsell, C. A., Weatherford, D., and Wixted, J. T. (2017) ROCs in Eyewitness Identification: Instructions versus Confidence Ratings. Appl. Cognit. Psychol., 31: 467– 477. https://doi.org/10.1002/acp.3344

Mickes, L., Flowe, H. D., & Wixted, J. T. (2012). Receiver operating characteristic analysis of eyewitness memory: Comparing the diagnostic accuracy of simultaneous and sequential lineups. Journal of Experimental Psychology: Applied, 18, 361–376. https://doi.org/10.1037/a0030609

Mickes, L., Moreland, M. B., Clark, S. E., & Wixted, J. T. (2014). Missing the information needed to perform ROC analysis? Then compute d′, not the diagnosticity ratio. Journal of Applied Research in Memory and Cognition, 3(2), 58–62. https://doi.org/10.1016/j.jarmac.2014.04.007

Mishra, M. V., Fry, R., Saad, E., Arizpe, J., Ohashi, Y., & DeGutis, J. (2020, August 6). Comparing the sensitivity of face matching assessments to detect face perception deficits. https://doi.org/10.31234/osf.io/68gbm

Moreland, M. B. (2015). Decision Processes in Eyewitness Identification. (Ph.D.). University of California, Riverside, Ann Arbor.

Morgan, C. A., Hazlett, G., Baranoski, M., Doran, A., Southwick, S., & Loftus, E. (2007). Accuracy of eyewitness identification is significantly associated with performance on a standardized test of face recognition. International Journal of Law & Psychiatry, 30, 213–223. https://doi.org/10.1016/j.ijlp.2007.03.005.

Murray, D. M., & Wells, G. L. (1982). Does Knowledge That a Crime Was Staged Affect Eyewitness Performance? Journal of Applied Social Psychology, 12(1), 42-53. https://doi.org/10.1111/j.1559-1816.1982.tb00847.x

Murray, E., Bennetts, R., Tree, J., & Bate, S. (2021). An Update of the Benton Facial Recognition Test. https://doi.org/10.31234/osf.io/6bt3z

National Institute of Justice (US). Technical Working Group for Eyewitness Evidence. (2003). Eyewitness evidence: A trainer's manual for law enforcement (Vol. 1). US Department of Justice, Office of Justice Programs, National Institute of Justice.

National Institute of Justice. (1999). Eyewitness evidence: A guide for law enforcement. Washington: DIANE Publishing.
National Research Council. (2014). Identifying the Culprit : Assessing Eyewitness Identification. Washington, DC: The National Academies Press. https://doi.org/10.17226/18891
National Research Council. (2014). Identifying the Culprit: Assessing Eyewitness Identification. Washington, DC: The National Academies Press. https://doi.org/10.17226/18891
Navon, D. (1992). Selection of lineup foils by similarity to suspect is likely to misfire. Law and Human Behavior, 16, 575–593. https://doi.org/10.1007/BF01044624
Newcombe, F. (1979). The processing of visual information in prosopagnosia and acquired dyslexia: Functional versus physiological interpretation. Research in psychology and medicine, 1, 315-322.
Nosworthy, G. J., & Lindsay, R. C. L. (1990). Does Nominal Lineup Size Matter? Journal of Applied Psychology, 75(3), 358-361. https://doi.org/10.1037/0021-9010.75.3.358
Nunn, J. A., Postma, P., & Pearson, R. (2001). Developmental prosopagnosia: Should it be taken at face value?. Neurocase, 7(1), 15-27. https://doi.org/10.1093/neucas/7.1.15
Oriet, C., & Fitzgerald, R. J. (2018). The Single Lineup Paradigm: A New Way to Manipulate Target Presence in Eyewitness Identification Experiments. Law & Human Behavior, 42(1), 1-12. https://doi.org/10.1037/lhb0000272**
Palmer, M. A., Brewer, N., Weber, N., & Nagesh, A. (2013). The confidence- accuracy relationship for eyewitness identification decisions: Effects of exposure duration, retention interval, and divided attention. Journal of Experimental Psychology: Applied, 19(1), 55–71. https://doi.org/10.1037/a0031602
Pell, M. D. (2006). Cerebral mechanisms for understanding emotional prosody in speech. Brain & Language, 96(2), 221-234.
Police and Criminal Evidence Act 1984, Code D, (2017) Retrieved from https://www.gov.uk/government/publications/pace-code-d-2017
Police Scotland. (2018). Identification procedures: Standard operating procedure. https://www.scotland.police.uk/assets/pdf/151934/184779/identification-procedures-sop
Public Health Resource Unit (2006). The Critical Skills Appraisal Programme: making sense of evidence. Public Health Resource Unit, England. Retrieved from: http://www.casp-uk.net/
Quigley-Quigley-McBride, A., & Wells, G. L. (2021). Methodological considerations in eyewitness identification experiments. In Methods, measures, and theories in eyewitness identification tasks (pp. 85-112). Routledge.
Ramon, M., Bobak, A. K., & White, D. (2019). Super-recognizers: From the lab to the world and back again. British journal of psychology, 110(3), 461-479. https://doi.org/10.1111/bjop.12368
Read, J. D., Tollestrup, P., Hammersley, R., McFadzen, E., & Chris- tensen, A. (1990). The unconscious transference effect: Are bystanders ever misidentified? Applied Cognitive Psychology, 4, 3–31. https://doi.org/10.1002/ acp.2350040103
Read, J.D., Lindsay, D.S., Nicholls, T. (1998). The relationship between accuracy and confidence in eyewitness identification experiments: Is the conclusion changing? In Thompson, C.P., Bruce, D., Read, J.D., Hermann, D., Payne, D., Toglia, M.P., (1998), Eyewitness memory: Theoretical and applied perspectives (pp. 107–130). Lawrence Erlbaum Associates Publishers.
Rigon, A., Voss, M. W., Turkstra, L. S., Mutlu, B., & Duff, M. C. (2018). Different aspects of facial affect recognition impairment following traumatic brain injury: The role of

perceptual and interpretative abilities. Journal of Clinical and Experimental Neuropsychology, 40(8), 805-819.

Roberts, R. J., & De Hamsher, K. S. (1984). Effects of minority status on facial recognition and naming performance. Journal of Clinical Psychology, 40(2), 539-545. https://doi.org/10.1002/1097-4679(198403)40:2%3C539::AID-JCLP2270400226%3E3.0.CO;2-8

Rondot, P., Tzavaras, A., & Garcin, R. (1967). Sur un cas de prosopagnosie persistant depuis quinze ans. Revue Neurologique, 117(3), 424-428.

Rossion, B., & Michel, C. (2018). Normative accuracy and response time data for the computerized Benton Facial Recognition Test (BFRT-c). Behavior Research Methods, 50(6), 2442–2460. https://doi.org/10.3758/s13428-018-1023-x

Rotello, C. M., & Chen, T. (2016). ROC curve analyses of eyewitness identification decisions: An analysis of the recent debate. Cognitive Research: Principles and Implications, 1(1), 1-12. https://doi.org/10.1186/s41235-016-0006-7

Russell, R., Duchaine, B., & Nakayama, K. (2009). Super-recognizers: People with extraordinary face recognition ability. Psychonomic bulletin & review, 16(2), 252-257. https://doi.org/10.3758/PBR.16.2.252

Searcy, J. H., Bartlett, J. C., & Memon, A. (1999). Age differences in accuracy and choosing in eyewitness identification and face recognition. Memory & cognition, 27(3), 538-552.https://doi.org/10.3758/BF03211547

Searcy, J. H., Bartlett, J. C., Memon, A., & Swanson, K. (2001). Aging and lineup performance at long retention intervals: Effects of metamemory and context reinstatement. Journal of Applied Psychology, 86(2), 207. https://psycnet.apa.org/doi/10.1037/0021-9010.86.2.207

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on generality (COG): A proposed addition to all empirical papers. Perspectives on Psychological Science, 12(6), 1123-1128. https://doi-org.bham-ezproxy.idm.oclc.org/10.1177/1745691617708630

Smalarz, L., & Wells, G. L. (2014). Post-identification feedback to eyewitnesses impairs evaluators' abilities to discriminate between accurate and mistaken testimony. Law and human behavior, 38, 194-202. https://doi.org/10.1037/lhb0000067

Smith, A. M., Smalarz, L., Wells, G. L., Lampinen, J. M., & Mackovichova, S. (2022, March 28). Fair Lineups Improve Outside Observers' Discriminability, Not Eyewitnesses' Discriminability: Evidence for Differential Filler-Siphoning Using Empirical Data and the WITNESS Computer-Simulation Architecture. Journal of Applied Research in Memory and Cognition. Advance online publication. http://dx.doi.org/10.1037/mac0000021**

Smith, A. M., Wells, G. L., Smalarz, L., & Lampinen, J. M. (2018). Increasing the similarity of lineup fillers to the suspect improves the applied value of lineups without improving memory performance: Commentary on Colloff, Wade, and Strange (2016). Psychological Science, 29(9), 1548-1551.

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

Starns, J., Cohen, A. L., & Tuttle, M. D. (2022). A theory-based approach for constructing recognition Receiver Operating Characteristics (ROCs) in complex tasks, with an application to full lineup ROCs. https://doi.org/10.31234/osf.io/5wp7c

Starzynski, L. L., Ullman, S. E., Filipas, H. H., & Townsend, S. M. (2005). Correlates of women's sexual assault disclosure to informal and formal support sources. Violence and Victims, 20, 417–432. https://doi.org/10.1891/0886-6708.20.4.417

Steblay, N. K., Dysart, J. E., & Wells, G. L. (2011). Seventy-two tests of the sequential lineup superiority effect: A meta-analysis and policy discussion. Psychology, Public Policy, and Law, 17(1), 99. https://doi.org/10.1037/a0021650

Steblay, N., Dysart, J., Fulero, S., & Lindsay, R. C. L. (2003). Eyewitness accuracy rates in police showup and lineup presentations: A meta-analytic comparison. Law and Human Behavior, 27(5), 523-540. https://doi.org/10.1023/A:1025438223608

Steblay, N.M. (1997). Social Influence in Eyewitness Recall: A Meta-Analytic Review of Line-up Instruction Effects. Law and Human Behaviour, 21 (3), 283-297. https://doi.org/10.1023/A:1024890732059

Trahan, D. E. (1997). Relationship between facial discrimination and visual neglect in patients with unilateral vascular lesions. Archives of Clinical Neuropsychology, 12(1), 57–62. https://doi.org/10.1093/arclin/12.1.57

Tredoux, C., Parker, J. F., & Nunez, D. (2007). Predicting eyewitness identification accuracy with mock witness measures of lineup fairness: Quality of encoding interacts with lineup format. South African Journal of Psychology, 37(2), 207-222. https://doi.org/10.1177/008124630703700201

Tunnicliff, J. L., & Clark, S. E. (2000). Selecting foils for identification lineups: Matching suspects or descriptions? Law and Human Behavior, 24, 231–258. https://doi.org/10.1023/A:1005463020252**

Tzavaras, A., Merienne, L., & Masure, M. C. (1973). Prosopagnosia, amnesia and language disorders caused by left temporal lobe injury in a left-handed man. L'encephale, 62(4), 382-394.

Valentine, T. (1991). A unified account of the effects of distinctiveness, inversion and race in face recognition. Quarterly Journal of Experimental Psychology, 43A, 161–204. doi:10.1080/14640749108400966

Wade, K. A., Nash, R. A., & Lindsay, D. S. (2018). Reasons to Doubt the Reliability of Eyewitness Memory: Commentary on Wixted, Mickes, & Fisher (2018). Perspectives on Psychological Science. 13, 339-342. https://doi.org/10.1177/1745691618758261

Wallis, K., Kelly, M., McRae, S. E., McDonald, S., & Campbell, L. E. (2021). Domains and measures of social cognition in acquired brain injury: A scoping review. Neuropsychological rehabilitation, 1-35. https://doi.org/10.1080/09602011.2021.1933087

Wang, R., Li, J., Fang, H., Tian, M., & Liu, J. (2012). Individual differences in holistic processing predict face recognition ability. Psychological science, 23(2), 169-177. https://doi.org/10.1177%2F0956797611420575

Wells, G. L. (1978). Applied eyewitness-testimony research: System variables and estimator variables. Journal of Personality and Social Psychology, 36(12), 1546.

Wells, G. L. (1993). What do we know about eyewitness identification? American Psychologist, 48,553–571.http://dx.doi.org/10.1037/0003-066X.48.5.553

Wells, G. L., & Bradfield, A. L. (1998). " Good, you identified the suspect": Feedback to eyewitnesses distorts their reports of the witnessing experience. Journal of Applied Psychology, 83, 360. https://doi.org/10.1037/0021-9010.83.3.360

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. https:// /doi.org /10.1037/lhb0000359

Wells, G. L., Leippe, M. R., & Ostrom, T. M. (1979). Guidelines for empirically assessing the fairness of a lineup. Law and Human Behavior, 3, 285–293. doi:10.1007/BF01039807

Wells, G. L., & Penrod, S. D. (2011). Eyewitness identification research: Strengths and weaknesses of alternative methods. *Research methods in forensic psychology*, 237-256.

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

Wells, G. L., 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, 313–317. doi:10.1016/j.jarmac.2015.08.008

Wells, G. L., Small, M., Penrod, S., Malpass, R. S., Fulero, S. M., & Brimacombe, C. A. E. (1998). Eyewitness identification procedures: Recommendations for lineups and photospreads. Law and Human Behavior, 22, 603–643.

Wetmore, S. A., Neuschatz, J. S., Gronlund, S. D., Wooten, A., Goodsell, C. A., & Carlson, C. A. (2015). Effect of retention interval on showup and lineup performance. Journal of Applied Research in Memory and Cognition, 4(1), 8-14. https://doi.org/10.1016/j.jarmac.2014.07.003 **

Wilmer, J. B., Germine, L., Chabris, C. F., Chatterjee, G., Gerbasi, M., & Nakayama, K. (2012). Capturing specific abilities as a window into human individuality: The example of face recognition. Cognitive neuropsychology, 29(5-6), 360-392. https://doi.org/10.1080/02643294.2012.753433

Wilmer, J. B., Germine, L., Chabris, C. F., Chatterjee, G., Williams, M., Loken, E., & Duchaine, B. (2010). Human face recognition ability is specific and highly heritable, Proceedings of the National Academy of Sciences, USA, 107 (pp. 5238–5241). https://doi.org/10.1073/pnas.0913053107.

Wilson, B. M., & Colloff, M. (2020). Coherently creating full receiver operating characteristic curves of police lineups. Poster presented at the 61st annual meeting of the Psychonomic Society (Virtual).

Wilson, B. M., Seale-Carlisle, T. M., & Mickes, L. (2018). The effects of verbal descriptions on performance in lineups and showups. Journal of Experimental Psychology: General, 147(1), 113–124. https://doi.org/10.1037/xge0000354

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

Wixted, J. T., & Mickes, L. (2018). Theoretical vs. empirical discriminability: the application of ROC methods to eyewitness identification. Cognitive Research: Principles and Implications, 3(1), 9. https://doi.org/10.1186/s41235-018-0093-8

Wixted, J. T., & Wells, G. L. (2017). The relationship between eyewitness confidence and identification accuracy: A new synthesis. Psychological Science in the Public Interest, 18, 10- 65. https://doi.org/10.1177/1529100616686966

Wogalter, M.S., Malpass, R.S., & McQuiston, D.E. (2004). A national survey of U.S. police on preparation and conduct of identification lineups. Psychology, Crime, & Law, 10, 69–82. https://doi.org/10.1080/10683160410001641873

Wood, S. (2017). Lineup Construction Fairness and Phenotypic Bias. (M.A.). City University of New York John Jay College of Criminal Justice, Ann Arbor.

Woodhead, M. M., & Baddeley, A. D. (1981). Individual differences and memory for faces, pictures, and words. Memory & Cognition, 9(4), 368-370. https://doi.org/10.3758/BF03197561

Yeates, G., Edwards, A., Murray, C., Creamer, N., & Mahadevan, M. (2013). The use of emotionally-focused couples therapy (EFT) for survivors of acquired brain injury with social cognition and executive functioning impairments, and their partners: a case series analysis. Neuro-Disability & Psychotherapy, 1(2), 151-194.

Yerys, B. E., Herrington, J. D., Bartley, G. K., Liu, H.-S., Detre, J. A., & Schultz, R. T. (2018). Arterial spin labeling provides a reliable neurobiological marker of autism spectrum disorder. Journal of Neurodevelopmental Disorders, 10 (1), 32. https://doi.org/10.1186/s11689-018-9250-0

Young, A. W., & Burton, A. M. (2017). Recognizing faces. Current Directions in Psychological Science, 26(3), 212-217. https://doi.org/10.1177%2F0963721416688114

Young, S. G., Hugenberg, K., Bernstein, M. J., & Sacco, D. F. (2012). Perception and Motivation in Face Recognition: A Critical Review of Theories of the Cross-Race Effect. Personality and Social Psychology Review, 16(2), 116–142. https://doi.org/10.1177/1088868311418987

Zarkadi, T., Wade, K. A., & Stewart, N. (2009). Creating fair lineups for suspects with distinctive features. Psychological Science, 20(12), 1448-1453. https://doi.org/10.1111%2Fj.1467-9280.2009.02463.x

Zhu, Q., Song, Y., Hu, S., Li, X., Tian, M., Zhen, Z., … Liu, J. (2010). Heritability of the specific cognitive ability of face perception. Current Biology, 20, 137–142. https://doi.org/10.1016/j.cub.2009.11.067.

Appendices

Appendix A - Systematic Literature Review Search Record

Database	Search Terms	Date of search and number of Hits		Date of search and number of hits	
SCOPUS	1. Witness* PRE/2 (identif* or accura* or confiden* or discriminability* or bias*) 2. Lineup PRE/3 (filler* or Foil* or Similar* or select* or match* or appear* or construct* or compos* or fair* or unfair*) 3. 1 AND 2	03.07.20	1. 1207 2. 731 3. 212	19.04.22	1. 1392 2. 847 3. 242
EBSCO HOST	1. Witness* n2 (identif* or accura* or confiden* or discriminability* or bias*) 2. Lineup n3 (filler* or Foil* or Similar* or select* or match* or appear* or construct* or compos* or fair* or unfair*) 3. 1 AND 2	05.07.20	1. 318 2. 95 3. 14	19.04.22	1.270 2. 80 3. 14
Web of Science	1. Witness* NEAR/2 (identif* or accura* or confiden* or discriminability* or bias*) 2. Lineup NEAR/3 (filler* or Foil* or Similar* or select* or match* or appear* or construct* or compos* or fair* or unfair*) 3. 1 AND 2	03.07.20	1. 616 2. 214 3. 32	19.04.22	1. 734 2. 258 3. 17
PUBMED	1. Witness* near/2 (identif* or accura* or confiden* or discriminability* or bias*) 2. Lineup near/3 (filler* or Foil* or Similar* or select* or match* or appear* or construct* or compos* or fair* or unfair*) 3. 1 AND 2	03.07.20	1. 37 2. 11 3. 3	19.04.22	1. 33071 2. 337 3. 25
ProQuest	1. Witness* near/2 (identif* or accura* or confiden* or discriminability* or bias*) 2. Lineup near/3 (filler* or Foil* or Similar* or select* or match* or appear* or construct* or compos* or fair* or unfair*) 3. 1 AND 2	05.07.20	1. 14743 2. 9745 3. 72	19.04.22	1. 14944 2. 10 376 3. 75

PsychINFO	1. Witness* adj2 (identif* or accura* or confiden* or discriminability* or bias*) 2. Lineup adj3 (filler* or Foil* or Similar* or select* or match* or appear* or construct* or compos* or fair* or unfair*) 3. 1 AND 2	03.07.20	1. 1165 2. 968 3. 373	19.04.22	1. 3094 2.1076 3. 698

Appendix B- Systematic Literature Review Quality Assessment Tool

Screening Questions

1. Did the study ask a clearly-focused question? ❑ Yes ❑ Can't tell ❑ No

Consider if the question is 'focused' in terms of:

– the population studied

– the intervention given

– the outcomes considered

2. Was this a randomised controlled trial (RCT) and was it appropriately so? ❑ Yes ❑ Can't tell ❑ No

Consider:

– why this study was carried out as an RCT

– if this was the right research approach for the question being asked

Is it worth continuing?

Detailed Questions

3. Were participants appropriately allocated to intervention and control groups? ❑ Yes ❑ Can't tell ❑ No

Consider:

– how participants were allocated to intervention and control groups. Was the process truly random?

– whether the method of allocation was described. Was a method used to balance the randomization, e.g. stratification?

– how the randomization schedule was generated and how a participant was allocated to a study group

– if the groups were well balanced. Are any differences between the groups at entry to the trial reported?

– if there were differences reported that might have explained any outcome(s) (confounding)

4. Were participants, staff and study personnel 'blind' to participants' study group? ❑ Yes ❑ Can't tell ❑ No

Consider:

– the fact that blinding is not always possible

– if every effort was made to achieve blinding

– if you think it matters in this study

– the fact that we are looking for 'observer bias'

5. Were all of the participants who entered the trial accounted for at its conclusion? ❑ Yes ❑ Can't tell ❑ No

Consider:

– if any intervention-group participants got a control-group option or vice versa

– if all participants were followed up in each study group (was there loss-to-follow-up?)

– if all the participants' outcomes were analysed by the groups to which they were originally allocated (intention-to-treat analysis)

– what additional information would you liked to have seen to make you feel better about this

6. Were the participants in all groups followed up and data collected in the same way? ❑ Yes ❑ Can't tell ❑ No

Consider:

– if, for example, they were reviewed at the same time intervals and if they received the same amount of attention from researchers and health workers. Any differences may introduce performance bias.

7. Did the study have enough participants to minimise the play of chance? ❑ Yes ❑ Can't tell ❑ No

Consider:

– if there is a power calculation. This will estimate how many participants are needed to be reasonably sure of finding something important (if it really exists and for a given level of uncertainty about the final result).

8. How are the results presented and what is the main result?

Consider:

- *if, for example, the results are presented as a proportion of people experiencing an outcome, such as risks, or as a measurement, such as mean or median differences, or as survival curves and hazards*
- *how large this size of result is and how meaningful it is*
- *how you would sum up the bottom-line result of the trial in one sentence*

9. How precise are these results?

Consider:

- *if the result is precise enough to make a decision*
- *if a confidence interval were reported. Would your decision about whether or not to use this intervention be the same at the upper confidence limit as at the lower confidence limit?*
- *if a p-value is reported where confidence intervals are unavailable*

10. Were all important outcomes considered so the results can be applied? ❑ Yes ❑ Can't tell ❑ No

Consider whether:

- *the people included in the trail could be different from your population in ways that would produce different results*
- *your local setting differs much from that of the trial*
- *you can provide the same treatment in your setting*

Consider outcomes from the point of view of the:

- *individual*
- *policy maker and professionals*
- *family/carers*
- *wider community*

Consider whether:

- *any benefit reported outweighs any harm and/or cost. If this information is not reported can it be filled in from elsewhere?*
- *policy or practice should change as a result of the evidence contained in this trial*

Appendix C – Systematic Literature Review Quality Assessment

Quality Assessment - Screening Questions

1. **Did the experiment ask a clearly-focussed question?**
2. **Was this a randomised controlled trial (RCT) and was it appropriately so?**
3. **Were participants appropriately allocated to intervention and control groups?**
4. **Were participants, staff, and experiment personnel 'blind' to participants' experiment group?**
5. **Were all of the participants who entered the trial accounted for at its conclusion?**
6. **Were the participants in all groups followed up and data collected in the same way?**
7. **Did the experiment have enough participants to minimise the play of chance?**
8. **How are the results presented and what is the main result?**
9. **How precise are these results?**
10. **Were all important outcomes considered so the results can be applied?**

**Note questions 8 and 9 were not used due to lack of quantitative measure on the quality assessment tool to compare experiments*

Quality Assessment - Screening Scores

Key: Yes = 2, Can't Tell= 1, No=0, Exclude experiments with a score of 7 or below

Author(s)	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q10	Total /16	Target absent (TA) / Target Present (TP)	Included or Excluded	Comments *(what similarity is manipulated, how was innocent suspect chosen, exclusion reason)*
Bergold & Heaton (2018)	2	2	1	2	2	2	1	2	14	TA+TP	I	Impact of database size on filler similarity and identification. Filler selected through database based on similarity to target. Innocent suspect selected using database and most similar to each target used.
Booth (2019)	/	/	/	/	/	/	/	/	/	/	E	Secondary analysis. Excluded.
Brewer & Wells (2006)	2	2	1	1	2	2	2	2	12	TA+TP	I	Foils and innocent target selected on basis of similarity to target description.
Carlson et al. (2008)	2	2	2	0	2	1	1	2	12	TA +TP	I	Innocent suspect selected by using face with highest similarity rating to target. Foils selected using database and similarity based on similarity to physical characteristics of target.
Charman et al. (2011)	2	2	1	1	2	2	1	2	13	TA only	E	Similarity manipulated within lineup, e.g., a lineup included "dud" fillers (2 good fillers (moderate similarity) and four bad fillers (low similarity). No designated innocent suspect. Excluded as TA only.
Clarke & Tunnicliff (2001)	2	1	2	1	2	1	1	2	12	TA+TP	E	Innocent suspect selected on match to suspect description Foil to suspect similarity rated. Excluded due to lack of reported data for analysis.
Cohen et al. (2020)	/	/	/	/	/	/	/	/	/	/	E	Secondary data, applying theoretical model. Excluded

Colloff et al. (2016)	2	2	2	2	2	2	2	2	16	TA+ TP	I	Foils similarity to modal descriptions of target. TA lineups used a foil face with distinctive feature added.
Colloff et al. (2017)	2	2	2	2	2	2	2	2	16	TA+TP	I	Filler similarity matched to description of culprit, no innocent suspect.
Colloff et al. (2021)	2	2	2	2	2	2	2	2	16	TA + TP	I	Median similarity innocent suspect selected, appearance and description similarity manipulation.
Cutler et al. (1987)	2	2	1	0	2	1	1	2	11	TA+TP	E	Foil similarity to target manipulated. No details of innocent suspect appears foils were used in TA condition. Excluded due to lack of reported data for analysis.
Darling et al. (2008)	2	1	1	0	2	1	2	2	11	TA+TP	E	Similarity manipulated by foil match to suspect description or resemblance to suspect. Innocent suspect chosen as someone who closely resembled culprit. Excluded due to lack of reported data for analysis.
Devenport & Cutler (2004)	2	2	1	0	0	0	1	0	6	TP	E	Excluded due to focus on influence of expert testimony and low quality score.
Fitzgerald et al.(2015)	2	1	1	2	2	2	1	2	13	TA+TP	I	Foil similarity to culprit manipulated. Innocent suspect selected and foils created based on similarity to innocent suspect.
Flowe & Ebbesen (2007)	2	2	2	1	2	2	1	2	14	TA + TP	I	Foil similarity to target, innocent suspect was a "look-a-like) Also used a "removal without replacement procedure" (target removed and not replaced).
Flowe et al. (2014)	2	2	2	2	2	2	2	2	16	TP+TA	E	Similarity of suspect and filler emotion manipulated. Same-gender fillers selected from database. Innocent suspect

												"randomly designated" from fillers in target absent lineups. Excluded as experimental manipulation outside of the area of review.
Gonzalez et al. (1995)	2	0	1	1	2	0	1	2	9	TP	E	Manipulated similarity to target (high and low), no innocent suspect. Excluded as TP only.
Grondlund et al. (2009)	2	2	2	2	2	2	2	2	16	TA+TP	I	Foil similarity and innocent suspect matched to description of perp (using Florida Supervised Offenders database).
Horry & Brewer (2016)	2	0	2	1	2	2	1	2	12	TA+TP	I	Foil similarity to target / innocent suspect manipulated, innocent suspect similar to target.
Juslin et el. (1996)	2	2	1	0	2	0	1	2	10	TA+TP	I	Foil similarity either by description or matched to features of target. TA foil was selected by match to culprit
Key et al. (2015)	2	2	2	2	2	2	2	2	16	TA+TP	I	Foil similarity and innocent suspect matched to description of perp (using Florida Supervised Offenders database).
Key et al. (2017)	2	2	2	2	2	2	1	2	15	TA+TP	I	Filler similarity to description of perpetrator, low similarity shared only one feature with perp. Innocent suspect was a photo rated most similar to suspect.
Levi (2016)	/	/	/	/	/	/	/	/	/	/	E	Exclude as it is just a summary of research, no experiment or analysis conducted.
Lindsay & Wells (1980)	2	2	1	0	2	1	1	2	11	TA+TP	I	Foil similarity to target manipulated. Innocent suspect chosen by match to target general description.
Lindsay et al. (1991)	2	2	2	1	2	2	1	2	14	TA+TP	I	Innocent suspect matched to target. Foil similarity to target manipulated. Excluded

												as manipulation outside the scope of this review.
Lindsay et al. (1994)	2	0	0	0	2	2	1	2	9	TP	E	Foil similarity: match to description and similarity to suspect (most similar). Excluded as TP only.
Lucas et al. (2021)	2	2	2	2	2	2	2	2	16	TA + TP	I	Match to description, innocent from high similarity filler pool.
Lucas & Brewer (2021)	2	2	2	2	2	2	2	2	16	TA + TP	I	Matched to suspect description and appearance, innocent suspect matched to description, moderately similar to perpetrator.
Moreland (2015)	2	2	2	2	2	2	2	2	16	TA+TP	I	Innocent suspect and foil similarity to perp manipulated. Expt 2&3 Foil "neighbour" similarity – foil next to suspect- was manipulated to be high or low.
Murray & Wells (1982)	2	2	2	2	2	0	1	2	13	TA + TP	E	Physical similarity of line-up members manipulated. Innocent suspect selected using similarity scores in pilot experiment. Excluded due to lack of reported data for analysis.
Nosworthy & Lindsay (1990)	2	1	1	0	2	0	1	2	9	TA+TP	E	Foil similarity to suspect manipulated, innocent suspect matched by similarity to target. Excluded as experimental manipulation did not fit scope of review.
Oriet & Fitzgerald (2018)	2	2	1	2	2	2	1	2	14	TA+TP	I	3 experiments manipulating filler similarity. Filler similarity to suspect manipulated. Innocent suspect selected by higher similarity to target.
Read et al. (1990)	/	/	/	/	/	/	/	/	/	/	E	Foil selection on similarity to description of target. Foil similarity not manipulated. Instead, a "bystander" similarity to target

												manipulated, as a "high similarity" and "low similarity".
Smith et al. (2020)	2	2	2	2	2	2	2	2	16	TA+TP	I	Suspect matched. Foils selected by match to description. Innocent suspect randomly selected from description matched filler pool.
Tredoux et al. (2009)	2	2	1	0	1	0	1	1	8	TA only	E	Foil similarity to target using match to description method. Innocent suspect chosen at random from pool of similar to target fillers. Excluded as TA only.
Tunnicliff & Clarke (2000)	2	1	1	2	2	1	1	2	12	TA+TP	I	Foils were matched to description in one condition and to description in another. Innocent suspect was chosen based on witness description.
Wetmore et al. (2015)	2	2	2	2	2	2	2	2	16	TA + TP	I	Foil similarity and innocent suspect matched to description of perp (using Florida Supervised Offenders database).
Wood (2017)	/	/	/	/	/	/	/	/	/	/	E	Master's thesis and focus on similarity by phenotype. Foil match to description and match to suspect.

Appendix D- Systematic Literature Review Data Extraction Form

Author(s)	Similarity Manipulation	Lineup Type	Lineup Medium	Were both TP and TA lineups used?	Were the same fillers used in TP and TA lineups?	How was the Innocent Suspect selected?	Is innocent suspect more similar to perp than other fillers in TA?	Were TA fillers matched to innocent suspect or the perpetrator? (appearance or description)	Any Other Relevant Information?	Notes for analysis (e.g collapse data?)

TP (Proportions)					TA (Proportions)				
n	Culpit ID Rate	Filler ID Rate	Reject Rate	Don't Know Rate	*n*	Innocent suspect ID Rate?	Filler ID Rate	Reject Rate	Don't Know Rate

(Data collected for lower higher; low, mid, higher; unfair, fair similarity comparisons)

Appendix E – Research Participant Information Sheet and Consent Form

UNIVERSITY OF BIRMINGHAM

INFORMATION SHEET

Title of the research project: Perception, memory, and decision-making

What is the purpose of the experiment?
The purpose of this experiment is to find out more about human perception, memory, and decision-making. Only people over the age of 16 have been invited to take part in this experiment.

Do I have to take part?
Participation is voluntary. We will ask you to tick a box to show you have agreed to take part. You are free to withdraw at any time by closing your internet browser, without giving a reason and without consequence. Should you wish to withdraw your data after completion you can email the researcher, within 72 hours of experiment completion, quoting your Mechanical Turk ID, Prolific ID or RPS ID (no reason needs be provided).

What will happen to me if I take part?
Before commencing the experiment, instructions will be given to the participant on setting up an adequate environment and on maintaining a comfortable posture during the experiment.
The experiment requires you to experiment lists of words, pictures, or short video clips of non-violent crimes on a computer screen. You will then be asked to make decisions about those pictures, words, or videos of non-violent crimes, or remember what those pictures, words or details about the video were. Participation will take approximately *10 minutes*.

What are the rewards for taking part?
If you sign up via the University of Birmingham Research Participation system you will be awarded *0.1-0.2 credits* for taking part. If you sign up via Amazon Mechanical Turk or Prolific, you will be paid in accordance with local norms (approximately $6.50 US per hour). If you sign up via a social media website, you will be entered into a prize draw to win a *£25 Amazon voucher*.

Are there any risks from taking part?
There are no risks, the material to be presented is mundane and not distressing.

What are the Covid-19 safety requirements?
To ensure Covid-19 safety, the activity can only take place using computer, tablet, or phone in your possession for the previous 72 hours. Participation should take place in a room that has been occupied only by members of the same social bubble for the previous 72 hours. Ensure that participation does not increase the chance to be in contact with individuals outside of your social bubble.

Will my taking part in the experiment be kept confidential?

Your data will not be associated with your name, only with a participant code (mturk ID/ Prolific/ RPS ID/ randomly generated ID). No personally identifying information will be shared or saved with your data. Data will be collected on Qualtrics' secure server. Once data collection is complete the data will be collated on the password protected computers / hard drives of the researchers and deleted from Qualtrics' servers. The data will be stored on the University of Birmingham servers - BEAR. If you signed up via a social media site, you will have the opportunity to leave your email address to be entered into the prize draw. Your email address will only be used by the researchers and will not be shared.

What will happen to the results of the research experiment?

The overall findings may be submitted for publication in a scientific journal or presented at scientific conferences. Following scientific publication, data will be aggregated (combined on spreadsheet) and made available to other researchers in aggregate form. Mechanical Turk user IDs, Prolific IDs and RSP ID numbers will be stripped from the data before it is shared. The data will be preserved and accessible for at least ten years. You will be able to obtain general information about the results of this research by contacting the researcher.

If you have questions or concerns about your rights as a research participant or about how the experiment is carried out, you may contact the lead researchers, Georgia Roughton [redacted] or Aleena Mahmood [redacted] in the first instance. However, if the query is not resolved, you may contact the project supervisor, Dr Melissa Colloff at [redacted]

Consent

1. I confirm that I have read and understood the experiment information
2. I confirm that I am over the age of 16
3. I understand that I can only withdraw my data within the first 72 hours
4. I agree that my data will be uploaded to a public repository after anonymization

By ticking this box, I confirm that I have read and understand the information about this experiment. ☐

Appendix F – Research Participant Debrief

PARTICPANT DEBRIEF

UNIVERSI
BIRMINGH

Thank you for participating. Your data will be stored on the password protected computers or the external hard drives of the researchers. Following scientific publication, all the data will be combined and made available to other researchers in combined form. Any personally identifying information (e.g., Mechanical Turk user IDs, Sona or Prolific RPS ID numbers) will be stripped from the data before it is shared. The data will be preserved and accessible for at least ten years.

The current experiment aims to investigate witness perception and memory for suspects with a distinctive facial feature. The findings of the experiment could have important implications for the way line ups are currently constructed in the United Kingdom to enhance witness identification performance while maintaining fairness of the lineup.

You are able to withdraw your data up to 72 hours from completing the experiment by emailing the lead researchers Aleena Mahmood [redacted] or Georgia Roughton [redacted] On the email, please quote your participant identification number, you do not need to provide a reason for your withdrawal and your data will be destroyed immediately. If you have any further questions or complaints in regard to your experience of the experiment please contact the lead researcher. If the query is not resolved, you may contact the project supervisor Dr Melissa Colloff [redacted]

Appendix G – Glossary

Term	Definition
Amazon Mechanical Turk	*This platform is used as a marketplace to match workers to available work. In chapter 3, this platform was used to access participants who completed the experiment for financial payment in accordance with local norms ($6.50 per 60 minutes).*
Benton Facial Recognition Test (BFRT)	*This is an assessment of unfamiliar facial recognition ability that was originally devised by Benton and Van Allen (1968).*
Block Lineup Construction	*This is a method used to create lineups for suspects with distinctive facial features. It involves covering the distinctive feature on the suspect with a black 'block' and then covering the corresponding area on the other lineup members with the same block (see Figure 5c).*
Ceiling affect	*This is when participants scores on a measure cluster towards the higher end of the measure (i.e., improved performance). In the context of the BFRT, this would suggest that the measure may not provide an accurate representation of facial recognition ability.*
Computerised Benton Facial Recognition Test (BFRTc)	*This is a computerised assessment of unfamiliar facial recognition ability by Rossion and Michel (2018). It has been developed from the long form Benton Facial Recognition Test and utilises the same method and test images. A key difference in this test is that participants are told to complete the task as fast as possible.*

Cross-race effect	*A phenomenon whereby individuals are better at recognising faces of their own race faces, than for other race faces.*
Designated Innocent Suspect	*A methodological characteristic of a lineup experimental study, whereby an individual is chosen to be the innocent suspect in all target absent lineups. This means that in every lineup where the guilty suspect is not present, then the designated innocent suspect will be present instead.*
Diagnostic Feature Detection Theory (DFD)	*This theory argues that when making a lineup identification decision, optimal witnesses will discount shared lineup member features that are non-diagnostic (i.e., features that are not indicative of guilt because they are shared across all lineup members) and instead focus on diagnostic features that are not shared across all lineup members (Wixted & Mickes, 2014).*
Diagnostic of Guilt	*Unique features of the guilty suspect, that are diagnostic of guilt because they are not shared by other lineup members.*
Discriminability D-Prime (d′),	*A measure of theoretical discriminability that can be calculated using the conceptual formula provided by Mickes et al. (2014); d' = z(correct ID rate) – z(false ID rate). A higher d' value indicates a better ability for the witness to discriminate between the innocent and guilty suspect and a d' value of 0 indicates an inability of witnesses to discriminate between innocent and guilty suspects (Macmillan & Creelman, 2004).*

Distinctive Suspect	*An individual who is thought to be guilty of a crime and has a distinctive facial feature such as scaring, a black eye or facial tattoos (see Figure 5a for an example).*
Do Nothing Lineup	*An unfair lineup in which the suspect stands out, i.e., the suspect (guilty or innocent) is the only person with a distinctive facial feature.*
Empirical Discriminability	*This refers to the degree to which a witness is able to accurately sort innocent and guilty suspects into their respective groups (Wixted & Mickes, 2018). It is calculated statistically by the area under the ROC curve (AUC) analysis.*
Eyewitness (also known as witness)	*An individual who observes something. In this context of this thesis, an eyewitness or witness is someone who has observed a crime.*
Fair Lineup	*A lineup where someone who had not seen the perpetrator would likely not be able to identify the suspect from the lineup at a rate higher than chance, because the suspect does not stand out as being different in physical appearance to the other lineup members.*
False Alarm Rate (FAR)	*The rate at which the innocent suspect is identified in lineups in which the guilty suspect is not present.*
Feature Discounting	*A process whereby the witness does not use a distinctive facial feature to make an identification decision, because all other lineup members have the same distinctive facial feature. This was the result of the high similarity replication*

lineups in Chapter 3, where all lineup members had the same tattoo.

Feature Glossing *A process whereby the witness uses the mere existence of a distinctive facial feature as an indicator of guilt regardless of the innocence or guilt of the lineup member. This was the result of the do nothing lineups in Chapter 3, where only the suspect had a distinctive facial feature.*

Feature Matching Model (FMM) *This model assumes that a face is defined by a number of features and that each facial feature has several possible settings (Colloff et al., 2021). For example, the feature of eye colour may have settings of brown, blue, hazel, grey and green. And, after witnessing a crime, the witness will have stored in memory the unique features of the perpetrator's face. When presented with a lineup in which the perpetrator is present, the encoded features of the perpetrator in the witnesses' memory will match those of the perpetrator presented in the lineup. However, an innocent suspect and fillers in a lineup, who are not guilty, will not possess the same matching features as they are unique to the perpetrator. In a description matched lineup, fillers are selected for the lineup on the basis that they match the witness description, and so some of the perpetrator's features will be shared by the fillers and innocent suspect, and these features will be non-diagnostic of guilt. However, the perpetrator will possess unique features that are not shared by the fillers or innocent suspect in the lineup (i.e., those not in their description), which are diagnostic of guilt and can be relied upon by the witness in making an identification decision. Therefore, lineup conditions which maximise the ability of the witness*

	to focus on facial features that are diagnostic of guilt to make an identification decision, will improve witness accuracy.
Feature Scrutinising	*A process whereby the witness scrutinises the distinctive facial features of the lineup members, and the guilty suspects' distinctive facial feature can be used as diagnostic of guilt because it is not shared by all other lineup members. This was the result of the low similarity replication lineups in Chapter 3, where only the guilty suspect had a tribal facial tattoo and the other lineup members had similar but non-identical facial tattoos.*
Filler	*Individuals presented within a lineup who are known to be innocent.*
Filler – Suspect Similarity	*The degree of physical similarity between the suspect (guilty or innocent) and the other lineup members who are known as fillers.*
Filler Similarity	*An overall term that can be used to refer to similarity relations between the suspect (guilty or innocent) and fillers, and of the similarity between the fillers themselves.*
Filler Siphoning	*This is a theory that states that fillers in a lineup protect the innocent suspect from identification because the fillers siphon choices away from the innocent suspect. That is, the witness picks the fillers instead of the innocent suspect (Smith et al., 2018; Wells et al., 2015)*

Guilty Suspect (also known as 'perpetrator')	*An individual who is guilty of a crime.*
High Similarity Replication Lineup	*A lineup in which the distinctive feature of the suspect is replicated across the other lineup members. In chapter 3, a target present lineup consisted of a guilty suspect with a tribal tattoo, and all other lineup members had the same tribal tattoo. In a target absent lineup, the innocent suspect had a star tattoo and all other lineup members had the same star tattoo.*
Higher (or high) Similarity Lineup	*Relative to other conditions within the experiment, the fillers are of higher similarity to the suspect, however the lineup remains fair.*
Hit Rate (HR)	*The rate at which the guilty suspect is identified in lineups in which the guilty suspect is present.*
Identification	*When a witness selects a lineup member as the person they believe to have committed a crime.*
Identification Accuracy	*This refers to the witness ability to correctly identify the guilty suspect when they are present within the lineup.*
Incorrect Rejection	*When a witness rejects a lineup that does contain the guilty suspect.*
Innocent Suspect	*An individual who is thought to be guilty of a crime, but who is in fact innocent (i.e., did not commit the crime). In practice, the guilt or innocence of a suspect is not known*

	but this can be manipulated so that the guilt or innocence of a suspect is known by the researcher.
Investigator Discriminability	*This refers to the ability of the investigator to use eyewitness identification to discriminate between innocent and guilty suspects (Smith et al., 2020; Smith et al., 2022).*
Lineup (also known as lineup identification)	*An identification procedure whereby a suspect (guilty or innocent) is presented alongside other people (known as fillers). The witness is then asked to identify if the perpetrator of the crime is present or not.*
Long Form Benton Facial Recognition Test (BFRT)	*This is the originally developed assessment of unfamiliar facial recognition ability by Benton and Van Allen (1968). It involves administration with a stimulus booklet and them matching different photographs of the same Caucasian face. There are a total of 54 scorable responses and this test requires twenty minutes to be administered (Benton & Van Allen, 1968).*
Lower (or low) Similarity Lineup	*Relative to other conditions within the experiment, fillers are less similar to the suspect (i.e., perpetrator or innocent suspect, depending on suspect matching procedure used), however the lineup appears fair (e.g., the suspect does not appear to stand out to someone who does not have a memory of the perpetrator).*
Low Similarity Replication Lineup	*A lineup in which the suspect's distinctive feature is not directly replicated across the other lineup members. Instead similar but non-identical distinctive features were replicated across the lineup members. In target present lineups, the guilty suspect had a tribal tattoo and all other*

	lineup members had similar but non-identical facial tattoos. In target absent lineups, the innocent suspect had a star tattoo, and all other lineup members had a similar but non identical facial tattoo.
Match to Appearance	*When other lineup members (fillers) are selected on the basis of their appearance relative to the suspect.*
Match to Description	*When other lineup members (fillers) are selected on the basis of their description relative to the suspect.*
Median Similarity Innocent Suspect	*An innocent suspect who was selected on the basis that they were the in the middle of the range of similarity from lowest to highest similarity of a pool of description matched fillers and the guilty suspect (see Chapter 3 for more details).*
Memory distribution	*A memory is the range of the memory signal for each particular stimulus. I.e. there will be a memory distribution for the range of the memory signal generated by an innocent suspect. When there is more overlap between memory distributions for stimuli, then it becomes harder to discriminate between those stimuli.* *For example when there is more overlap between the memory distributions for the guilty perpetrator and the innocent suspect (such as due to the innocent suspect being highly similar in appearance to the guilty perpetrator) then it becomes harder for the witness to discriminate between the guilty perpetrator and innocent suspect, resulting in a decrease in discriminability.* *However, when there is less overlap between the memory distributions for the guilty perpetrator and the innocent suspect (i.e. as the innocent suspect is of moderate*

	similarity to the guilty perpetrator) then it becomes easier for the witness to discriminate between the guilty perpetrator and innocent suspect, resulting in an increase in discriminability).
Methodological Characteristics	*This refers to the way in which a lineup experiment has been conducted i.e., how fillers were matched to the suspect (description or appearance); how the innocent suspect was selected (highly similar, moderate, or low similarity) or how a lineup was presented (simultaneous or sequential).*
Moderate Similarity Lineup	*Relative to other conditions within the experiment, fillers are of moderate similarity to the suspect, however the lineup appears fair.*
Non-Diagnostic of Guilt	*Features that are not indicative of guilt because they are shared across all lineup members.*
Optimal Filler Similarity	*This refers to the lineup condition that results in improved ability for the witness to discriminate between innocent and guilty suspects, resulting in an increase in identifications of the guilty suspect, while protecting the innocent suspect from misidentification.*
Partial Area Under the Curve (pAUC)	*This is a measure of empirical discriminability. Specifically, partial area under the curve is an analysis where only the area of the receiver operator characteristic in which data have been observed are analysed. In Chapter 3, this meant that partial area under the curve analysis was completed on the basis of the smallest false alarm rate observed (see Chapter 3 for more details).*

Perpetrator matched	*When the other lineup members are selected on the basis of their match to the perpetrator. In target present lineups, this means matching the other lineup members to the guilty suspect. In target absent lineups, this means matching the other lineup members to the perpetrator, even though the innocent suspect is presented instead of the perpetrator.*
Pixelation Lineup Construction	*This is a method used to create lineups for suspects with distinctive facial features. It involves covering the distinctive feature on the suspect with an area of pixelation and then covering the corresponding area on the other lineup members with an area of pixelation (see Figure 5b).*
Pre-registration	*This refers to the practice of registering proposed hypotheses, methods, and analysis of an experiment before it has been conducted. The details of the experiment in chapter 3 and planned analysis were pre-registered (https://osf.io).*
Prosopagnosia	*This is a neurological disorder that occurs when an individual has an inability to recognise faces.*
Psychometric Tools	*These are tests used in psychological assessment that allow for an objective measures of psychological characteristics such as psychological symptoms, personality traits and mental capabilities. In chapter 4, a critique of a psychometric tool known as the Benton Facial Recognition Test (BFRT, Benton & Van Allen, 1968) is conducted.*
Qualtrics	*An online survey tool used to run the experiment in chapter 3.*

Receiver Operator Characteristic (ROC)	*Empirical discriminability (i.e., requiring no assumptions about underlying memory strength distributions) can be measured by Receiver Operating Characteristic (ROC) analysis. A ROC plot depicts the hit rate (perpetrator identifications) and false alarm rate (innocent suspect identifications). Lineup conditions yielding higher discriminability result in an increased hit rate of correct perpetrator identifications, and a decreased false alarm rate of innocent suspect identifications and are depicted in a higher ROC curve as shown in figure 1b. As also evident in figure 1b, higher ROC curves (an empirical measure of discriminability) typically equate to larger d' values (a measure of theoretical discriminability; see Mickes et al., 2014).*
Replication Lineup Construction	*This is a method used to create lineups for suspects with distinctive facial features. It involves replicating a suspects' distinctive feature across the other lineup members. There is variation in the application of this method in practice. It could involve a high similarity replication lineup, whereby the suspects' distinctive feature is replicated across all of the other lineup members (see Figure 6b). Or it could involve a low similarity replication lineup, whereby a similar but non-identical distinctive feature is replicated across all of the other lineup members (see Figure 6c). This is the subject of the research conducted in chapter 3.*
Revised Benton Facial Recognition Test (BFRTr)	*This is a revised computerised assessment of unfamiliar facial recognition ability by Murray et al. (2021) It has been developed from the long form Benton Facial Recognition Test and utilises the same method. A key difference in this test is that different test images were used.*

	Specifically, male only images were used, and the images were said to be more naturalistic as they included varied images for each male that were taken on different dates within a one year period, whereby hairstyle, skin tone, blemishes and lighting varied.
Sensitivity	*This refers to a tests' ability to identify a positive result. In the context of chapter 4, this would refer to the ability of the Benton Facial Recognition Test to identify a deficit in unfamiliar facial recognition.*
Sequential Lineup	*A lineup in which all lineup members are presented one at a time.*
Short Form Benton Facial Recognition Test (BFRT)	*This test is based on the long form assessment of unfamiliar facial recognition ability by Benton and Van Allen (1968). It also involves administration with a stimulus booklet and them matching different photographs of the same Caucasian face. This test differs from the long form Benton Facial Recognition Test as there are a total of 27 scorable responses and this test requires seven minutes to be administered (Benton et al., 1994).*
Showup	*An identification procedure in which the suspect is presented alone to the witness. The witness is then asked to identify if the suspect in the showup is guilty of the crime.*
Signal Detection Theory (SDT)	*This is concerned with a person's ability to discriminate the presence or absence of a stimulus.*
Simultaneous Lineup	*A lineup in which all lineup members are presented at the same time.*

Suspect	*An individual who is thought to be guilty of a crime.*
Suspect Matched	*When the other lineup members are selected on the basis of their match to the suspect. In target present lineups, this means matching the other lineup members to the guilty suspect. In target absent lineups, this means matching the other lineup members to the innocent suspect.*
Target Absent (TA)	*A lineup in which the guilty suspect is not present.*
Target Present (TP)	*A lineup in which the guilty suspect is present.*
Theoretical Discriminability	*This can be measured using the d-prime statistic (d') and refers to the amount of theoretical overlap between the memory strengths for innocent and guilty suspects in the witness's memory (Wixted & Mickes, 2018).*
Unfair Lineup	*A lineup where someone who had not seen the perpetrator would be able to identify the suspect from the lineup on the basis that the suspect stands out as being different in physical appearance to the other lineup members.*
Unique features	*Characteristics of the face of the guilty suspect that are not shared by the other lineup members.*
Z-tests	*This is a statistical test used to examine if two population means are different.*