

Lineup Construction Methods: Should Lineup Constructors be Blinded to the Suspect?

Courtlyn Elkins

Supervisors: Edwin Burns, Ruth Horry, and Hayley Young

Submitted to Swansea University in fulfilment of the requirements for
the Degree of Doctor of Philosophy

Swansea University 2026

Declarations

This work has not previously been accepted in substance for any degree and is not being concurrently submitted in candidature for any degree.

Signed _____

Date 02/06/2026

This thesis is the result of my own investigations, except where otherwise stated. Other sources are acknowledged by footnotes giving explicit references. A bibliography is appended.

Signed _____

Date 02/06/2026

I hereby give consent for my thesis, if accepted, to be available for photocopying and for interlibrary loan, and for the title and summary to be made available to outside organisations.

Signed _____

Date 02/06/2026

The University's ethical procedures have been followed and, where appropriate, that ethical approval has been granted.

Signed _____

Date 02/06/2026

Abstract

Lineups should contain known-innocent individuals (fillers) who resemble the suspect. Two common filler selection methods are: the *match-to-suspect method*, fillers are selected based on perceived similarity to the suspect; and the *match-to-description method*, fillers are selected based on the perpetrator description (Wells et al., 2020). In a systematic review, I examine how the match-to-description method has been implemented and which method leads to superior lineup outcomes.

This thesis aimed to investigate: (1) how filler selection methods impact lineup outcomes; (2) how blinding match-to-description lineup constructors impacts the method's effectiveness. This thesis presents four empirical studies in which stimuli were created (Study 1), and the resulting lineups were tested for fairness (Study 2), suspect-filler similarity and description-fit (Study 3), and eyewitness discriminability (Study 4).

Study 1, perpetrator descriptions were elicited from eyewitnesses using a free recall questionnaire or modified version of the Self-Administered Interview. Lineup constructors were provided with: (1) a suspect's image (*match-to-suspect* condition), (2) a perpetrator description (blinded *match-to-description* condition), or (3) a suspect's image and matching perpetrator description, with instructions to use the description (unblinded *match-to-description* condition).

Study 2 tested the lineups for fairness using the mock witness paradigm. Blinded match-to-description lineups yielded fairer lineups than match-to-suspect lineups. Blinding had limited impact on the fairness of match-to-description lineups.

Study 3 investigated suspect-filler similarity and lineup member description-fit. Match-to-suspect lineups yielded higher suspect-filler similarity and lower description-fit than blinded

match-to-description lineups. Blinding had limited impact on either metric for match-to-description lineups.

Study 4 tested the lineups in a standard eyewitness identification task. The blinded match-to-description method outperformed the match-to-suspect and unblinded match-to-description methods in discriminability and resultant fairness. Offering support for the match-to-description method over the match-to-suspect method and demonstrating the importance of blinding in lineup procedures. Thus, having real-world implications for improving police procedures regarding suspect lineup construction.

Acknowledgements

I would like to thank my mentors, family, and friends for supporting me through this experience. First, I would especially like to thank my supervisor, Dr. Ruth Horry for everything she has done for me throughout this PhD. From assisting with grant proposals to obtain funding, to being a constant source of support and guidance over the past three years, this PhD would not have been possible without you. I would also like to thank Dr. Edwin Burns for joining my supervision team and providing a new perspective and invaluable guidance over the past year.

Next, I would like to thank my friends for without you Swansea never would have become a home away from home. I truly do not know what I would do without each of you. From long days stuck in the office to sea dips and beach BBQs, thank each of you for keeping me grounded throughout this PhD. You all have truly become a second family to me over the years; I am forever grateful Swansea University brought all of us together.

Lastly, I would like to thank my dad and sisters for being so supportive of me pursuing my PhD so far from home. I wouldn't have made it these past few years without your words of encouragement and love. I am so lucky to have family who encourage me to step out of my comfort zone and push me to succeed. And to my dog, Princess, who was snuggled up by my side during the writing of this thesis – you made the long days of proof reading and edits worth it.

Mom – I hope you are proud. Thank you for teaching me how to be brave and how to believe in myself. If not for you, this PhD never would have been possible.

Table of Contents

List of Tables and Figures	viii
Chapter 1: Systematic Review and Meta-Analysis	1
Introduction	1
Systematic Review Method	15
Systematic Review Results	19
Systematic Review Discussion	26
Meta-Analysis	28
Meta-Analysis Method	28
Meta-Analysis Results	31
Meta-Analysis Discussion	45
Chapter 2: Eliciting Person Descriptions and Constructing Suspect Lineups	49
Introduction	49
Study 1a Method	63
Study 1a Results	66
Study 1a Discussion	71
Study 1b Method	75
Study 1b Results	78
Study 1b Discussion	82
General Discussion	83
Chapter 3: Filler Selection Strategies and Lineup Fairness	86
Introduction	86
Study 2 Method	93

Study 2 Results	95
Study 2 Discussion	107
Chapter 4: Suspect-Filler Similarity and Lineup Member Description-Fit	111
Introduction	111
Study 3a Method	122
Study 3a Results	124
Study 3a Discussion	129
Study 3b Method	131
Study 3b Results	132
Study 3b Discussion	136
General Discussion	136
Chapter 5: Filler Selection Strategies and Eyewitness Identification Decisions	141
Introduction	141
Study 4 Method	152
Study 4 Results	155
Study 4 Discussion	168
Chapter 6: General Discussion	174
References	199
Appendices	213

List of Tables and Figures

Tables

Table 1. Match-to-Description Implementations	20
Table 2. Risk of Bias Criteria (adapted from Horry et al., 2021)	24
Table 3. Risk of Bias Scores for Each Study Included in Systematic Review	25
Table 4. Test of Between-Subgroup Homogeneity for Correct IDs in Target Present Lineups	40
Table 5. Test of Between-Subgroup Homogeneity for Correct IDs in Target Present Lineups	41
Table 6. Test of Between-Subgroup Homogeneity for Filler IDs in Target Present Lineups	42
Table 7. Test of Between-Subgroup Homogeneity for Filler IDs in Target Present Lineups	42
Table 8. Discriminability Statistics for MTS and MTD Lineups	45
Table 9. Mean Number of Details in Person Descriptions by Stimulus and Quality Condition	68
Table 10. Descriptions Elicited by Description Quality Condition	70
Table 11. Self-Reported Selection Strategies of Unblinded MTD Lineup Constructors	81
Table 12. Mock Witness Lineup Fairness Measures (adapted from Mansour et al., 2017)	89
Table 13. Mean Proportion and Standard Deviation of Suspect Chosen by Stimulus	97
Table 14. Descriptive Statistics for Construction Method, Collapsed Across Target	

Presence	99
Table 15. Descriptive Statistics for Construction Method Across Target Present	
Lineups	100
Table 16. Descriptive Statistics for Construction Method Across Target Absent	
Lineups	100
Table 17. Lineup Fairness Values for Each of the 24 Lineups	
	101
Table 18. Mean Proportion and Standard Deviation for Decision Strategy Usage	
	106
Table 19. Mean Filler-Suspect Similarity Ratings by Lineup Construction Method ...	
	125
Table 20. Mean Suspect-Filler Similarity Scores Across Construction Method	
	128
Table 21. Identification Rates per Lineup Construction Method and Target	
Presence	158
Table 22. Mean Resultant Tredoux's E and Effective Size Values for Each Lineup	
Construction Method	166
Table 23. Mean Proportion Suspect Selections, Resultant Suspect Bias, and Resultant	
Defendant Bias Values for Each Lineup Construction Method	168

Figures

Figure 1. PRISMA Flow-Chart	18
Figure 2. Forest Plot Depicting Correct Identifications for Target Present Lineups	
	32
Figure 3. Forest Plot Depicting Innocent Filler Identifications for Target Present	
Lineups	34
Figure 4. Forest Plot Depicting Incorrect Non-Identifications for Target Present	
Lineups	35
Figure 5. Forest Plot Depicting Innocent Suspect Identifications for Target Absent	

Lineups	37
Figure 6. Forest Plot Depicting Innocent Filler Identifications for Target Absent	
Lineups	38
Figure 7. Forest Plot Depicting Correct Non-Identifications for Target Absent	
Lineups	39
Figure 8. Forest Plot Depicting G-Statistics and Weighted Mean	43
Figure 9. Lineup Fairness Measures by Construction Method and Stimulus for Target Present Lineups	103
Figure 10. Lineup Fairness Measures by Construction Method and Stimulus for Target Absent Lineups	104
Figure 11. Bar Graph Depicting the Lineup Construction Method and Target Presence Interaction for Suspect-Filler Similarity Ratings	127
Figure 12. Trial Process Flowchart	155
Figure 13. Interaction Between Lineup Construction Method and Target Presence for Filler IDs	160
Figure 14. ROC Curves for Each Lineup Construction Method Collapsed Across Stimuli	162
Figure 15. CAC for Each Lineup Construction Method Collapsed Across Stimuli	164
Figure 16. Boxplot Depicting the Mean Resultant Tredoux's E Values for Each Lineup Construction Method	167
Figure 17. Boxplot Depicting the Mean Resultant Effective Size Values for Each Lineup Construction Method	167

Chapter 1

A systematic review and meta-analysis: The influence of lineup construction on eyewitness identification reliability

Eyewitness identification continues to be used in various aspects of criminal investigations even though countless studies have shown this type of evidence can be flawed. Although research has shown eyewitness memory to be fallible, eyewitness identification evidence remains one of the most persuasive types of evidence and has contributed to more wrongful convictions than any other type of evidence (Innocence Project, 2024; Wells et al., 1998; Wells et al., 2020). To improve this form of evidence, much research has been conducted on the factors that affect eyewitness memory. Generally, these factors are broken down into two categories: estimator variables and system variables (Innocence Project, 2024; Wells & Olson, 2003). Estimator variables refer to those factors outside the control of the criminal justice system, including specific eyewitness characteristics such as age, visual acuity, quality of view (e.g. distance, lighting, obstructions), and the level of arousal experienced during the event. System variables, on the other hand, are those that are within the control of the criminal justice system including policy, procedure, and application of those procedures surrounding the construction (e.g. number of lineup members or selection of lineup members) and administration (e.g. simultaneous vs sequential presentation) of suspect lineups. While both estimator and system variables can play a role in eyewitness misidentifications, research has shown that system variables have played a role in many cases where an individual was wrongly convicted due to witness misidentifications (Innocence Project, 2024). Additionally, as estimator variables are outside of the control of the legal system, the primary way to reduce misidentifications is by changing the procedures regarding eyewitness identifications. Thus, further research on system

variables, such as suspect lineup construction, is needed in order to reduce the number of misidentifications.

Decades of research have culminated in a series of best-practice guidelines published in a white paper in 2020 (Wells et al., 2020). According to this research, suspect lineups should consist of one suspected individual and *at least* five known-innocent individuals, often referred as fillers or foils, and may be presented in various formats such as live, in-person, photographic, and more recently video (Wells et al., 2020). Additionally, to minimise bias, specific pre-lineup instructions should be given, informing the witness that it is possible culprit is not present in the lineup. By stating the individual may be absent from the lineup, it opens the witness to the possibility the culprit is not there and as a result helps reduce the pressure on the witness to make an identification, thereby reducing the likelihood of the witness guessing if they do not see the individual they remember (Wells et al., 2020). Crucially, it is widely agreed that the individual administering the lineup should be blind to the identity of the suspect. This means the administrator has no knowledge of who the police suspect is to minimize the chance of influencing the witness' decision, either intentionally or unintentionally. These procedures, along with the use of fillers in a lineup, help reduce witness misidentifications.

While using fillers in a lineup is crucial for reducing witness misidentifications, the fillers must be what is often referred to as 'quality' fillers; that is, the fillers must prevent the suspect from standing out. To achieve this standard, it is often recommended that the fillers closely match the culprit description provided by the witness or closely resemble the suspect (Wells et al., 2020). The use of these quality fillers helps by allowing for known errors to help reduce wrongful identifications, controlling for guessing by witnesses who are uncertain, reducing the likelihood of the witness using relative judgments to make their selections, and ensuring the

lineups are testing recognition memory rather than simple recall (Luus & Wells, 1991). While there is agreement that quality fillers are needed for a lineup to be considered fair, there is conflicting information on how those fillers should be selected. The two commonly used methods of filler selection are the *match-to-suspect method* and the *match-to-description method*.

The most widely used approach is the match-to-suspect (MTS) method of filler selection. This method is one in which the known-innocent fillers are selected based on their perceived physical similarity to the suspect (Malpass, Tredoux, & McQuiston-Surrett, 2007). This method follows the guidelines set in the Police and Criminal Evidence Act 1984, Code D, (PACE) which states that fillers should resemble the suspect based on physical characteristics such as race, gender, approximate age, build, and hair colour (Police and Criminal Evidence Act 1984, Code D, 2023). The other commonly used approach to selecting fillers is the match-to-description (MTD) method. This method is one in which the known-innocent fillers are selected based on how closely their physical characteristics match those mentioned in the culprit description provided by the witness (Luus & Wells, 1991). This method helps ensure that the fillers selected match the culprit seen by the witness. While each method in theory creates fair lineups, there has been much debate in the literature determining which method consistently yields the fairest lineups.

For instance, several studies have found that the MTD method leads to better identification outcomes than the MTS method. Wells and colleagues (1993) directly compared the two lineup construction methods in a study investigating the impact of lineup construction method on lineup outcomes. In this study, participants witnessed a staged crime and then were asked to give a written description of the perpetrator. These descriptions were used by MTD

lineup constructors to create lineups. In this study, MTD lineups were created by selecting the five fillers who best matched the perpetrator description provided by eyewitnesses, while MTS lineups were created by selecting five fillers who most closely resembled the suspect. These lineups were also compared to a third condition, the 'mismatch-description' method, where fillers were selected for purposely violating the perpetrator description on at least one major feature (e.g. hair colour or hair style) (Wells et al., 1993). Once these lineups had been created, witnesses were asked to view a lineup and to make an identification. They found that lineups created using the MTD method yielded higher correct identification rates compared to lineups created using the MTS method. Additionally, they found the MTD method was just as effective in reducing innocent suspect identifications as the MTS method. These findings suggest the MTD method may increase correct suspect identifications, while having no negative impact on innocent suspect identifications.

These findings were later replicated by Juslin and colleagues (1996) using active police officers as lineup constructors. In this study, police officers using the MTS method were given an image of the suspect, guilty or innocent, and asked to select seven fillers who resembled the suspect. Police officers using the MTD method were provided with a brief description of the perpetrator and asked to select seven fillers who best matched the information provided. They found that when the target was present in the lineup, the MTD method led to more correct identifications compared to the MTS method. They also found identical false identification rates, or innocent suspect identifications, for the two lineup construction methods when the target was absent from the lineup. These findings were further supported by research conducted by Carlson and colleagues (2019).

In a study directly comparing the MTD method to the MTS method, Carlson and

colleagues (2019) also found evidence supporting the use of the MTD method over the MTS method of lineup construction. For generalisability purposes, Carlson et al. (2019) compared several MTD and MTS lineups to one another. The MTD lineups were created by providing the description to four individuals who had no prior knowledge of the suspect, and instructing each of them to select 20 individuals from a large face database using the description provided. The researchers then randomly selected 10 fillers from the description-matched pool to serve as the fillers in the two description-matched target present lineups. This process was then repeated to create the target absent MTD lineups. The creation of the MTS lineups followed a similar process but instead used an image of the suspect to narrow the filler pool. The results of this study showed lineups created using the MTD method led to higher correct identification rates overall compared to lineups created using the MTS method. They also found no difference in false identification rates between the two lineup construction methods. Thus, these findings offer support for the MTD method of lineup construction.

While much of the existing research appears to show support for the MTD method, some studies have found that the method may lead to an increase in eyewitness misidentifications. In efforts to replicate the findings of Wells et al. (1993), Lindsay and colleagues (1994) tested two versions of the MTD method against the MTS method of lineup construction. However, one important difference between the studies is the use of similarity ratings by Lindsay et al. (1994) during the filler selection process. For MTS lineups, the five individuals who were rated highest for suspect-filler similarity were selected to be fillers. For the 'reasonable' MTD lineups, a filler was selected from the middle of each grouping of faces ranked by similarity ratings. This allowed for substantial variability in appearance among lineup members, as well as variability in suspect-filler similarity within the lineup. For the biased MTD lineups, the five foils who were

rated as least similar to the suspect, but still fitting the perpetrator description, were selected. In line with previous findings, Lindsay and colleagues (1994) found that lineups created using either version of the MTD method had higher correct identification rates compared to those created using the MTS method. However, when the target was absent from the lineup, they found the biased MTD method led to higher rates of innocent suspect identifications compared to the other two methods. They also found the MTS method led to the lowest rate of innocent suspect identifications. Thus, Lindsay and colleagues (1994) cautioned that while MTD lineups may be beneficial when the guilty perpetrator is present in the lineup, they may lead to an increase in eyewitness misidentifications when the perpetrator is absent.

This finding was supported by Jones (2021) in a study investigating the impact of filler selection method on lineup fairness and eyewitness identification decisions. This study used a different lineup construction methodology to compare four MTD lineups to a single MTS lineup. In the MTD lineups, all fillers were selected based on how well they matched specific features mentioned in the description. For example, 'GeneralCharacteristics' MTD lineups were created by ensuring fillers matched four general characteristics from the perpetrator description, while fillers in '4Internal' MTD lineups were selected for matching four internal features (e.g. eye colour/shape, nose, mouth) from the description. The process for selecting fillers for '4External' MTD lineups and '2Internal2External' MTD lineups was similar to the other two, the only difference being the features that were matched. Fillers in the MTS lineup were selected based on how closely they resembled the suspect. The results of this study showed no difference in correct identification rates between MTD and MTS lineups. This findings suggests there may not be an advantage for one method over the other in terms of increasing correct identifications. With that said, Jones (2021) also found that MTD lineups led to an increase in innocent suspect

identifications compared to MTS lineups. This finding further suggests MTD lineups may not provide adequate protection for innocent suspects against eyewitness misidentifications.

However, with that said, other research has found little to no difference between the two methods regarding lineup outcomes both correct identifications and false identifications.

Tunnicliff and Clark (2000) investigated the impact of filler selection on eyewitness identifications by directly comparing MTD lineups to MTS lineups. This study followed a lineup construction methodology similar to that of Juslin et al. (1996). For MTD lineups, detectives were provided a description of the perpetrator and asked to select five fillers who best matched the description. For MTS lineups, detectives were provided an image of the suspect and asked to select five fillers who most closely resembled the suspect. The results of this study showed no difference in correct identifications or innocent suspect identifications between the two methods. This suggests both lineup construction methods can lead to positive outcomes.

This finding was later supported by Darling and colleagues (2008). The lineup construction methodology used in this study was similar to that of Tunnicliff and Clark (2000) with the only difference being Darling et al. (2008) did not use detectives to construct the lineups. The MTD lineups were created by selecting the eight fillers who best matched the perpetrator description provided, while the MTS lineups were created by selecting the eight fillers who most closely resembled the suspect. The results of this study showed no difference in suspect identifications, both correct and innocent suspect, between the two lineup construction methods. This finding further suggests quality lineups can be created using both lineup construction methods. While the existing research is conflicted as to which method leads to better lineup outcomes, understanding the mechanisms behind these two lineup construction methods may provide useful insight into how they can protect against eyewitness

misidentifications.

Both the MTS method and the MTD method have their advantages and disadvantages. With the MTS method, matching the known-innocent fillers to the suspect helps prevent the suspect from standing out in the lineup, which is the main benefit to using this method. A second benefit to matching fillers to the suspect is that this method can be utilized when the culprit description provided by the witness is too vague or is lacking sufficient detail (Lindsay, Martin, & Webber, 1994). By selecting fillers who have similar physical characteristics to the suspect, lineup constructors can ensure the suspect does not stand out. However, the problem with this method arises with there being no guidance on where the line between similar enough and too similar is. Thus, this method can lead to lineups in which the fillers have high similarity to the suspect and one another. Wells and colleagues (1993) made a good argument against the MTS method when they suggested lineups created with this method can result in a lineup filled with clones, or identical individuals. No matter how good one's memory is, this task would be impossibly hard. This is because as similarity among lineup members increases, discriminability decreases, thereby decreasing the reliability and predictive utility of identification decision (Carlson et al., 2019). The importance of variability among lineup members for accurate witness decisions can be explained by the diagnostic feature detection theory.

The diagnostic feature detection (DFD) theory is one of perceptual learning which states that an important aspect of learning to discriminate between similar objects is through detecting the diagnostic features between them and being able to discount any non-diagnostic features. This then allows the decision maker to place more weight on any diagnostic features present to aid in their decisions. Therefore, discriminability increases when similar objects have existing diagnostic features that allow an individual to distinguish between them (Wixted & Mickes,

2014). When individuals in a lineup look too similar, there are fewer diagnostic features present to allow witnesses to accurately distinguish between guilty and innocent individuals. Thus, this lack of diagnostic features reduces a witness's ability to accurately identify the guilty perpetrator and can lead to false identifications. This is why the MTD method was proposed, as it has a natural stopping point (i.e. the features described in the perpetrator description) in characteristics the fillers should have. This natural stopping point allows for the lineup members to look similar to one another but still allows for more diagnostic features to be present among lineup members, thus making the MTD method a promising alternative to the MTS method of selecting lineup fillers.

With that said, Carlson and colleagues (2019) found that the MTD method of filler selection yields lineups with increased discriminability and higher amounts of “propitious heterogeneity” compared to the MTS method. Through the use of Receiver Operating Characteristic (ROC) analysis, this study found description-matched lineups led to increased empirical discriminability compared to suspect-matched lineups. Additionally, the results of this study showed none of the suspect-matched lineups led to better discriminability compared to description-matched lineups (Carlson et al., 2019). This finding suggests that the MTD method was more likely to lead to heterogeneous lineups which allows for the fillers and suspects to look similar to one another but still also allows for discriminability among lineup members (Carlson et al., 2019; Wixted & Mickes, 2014). This helps the witness distinguish between the innocent and guilty, as well as be more accurate in their decisions. This finding suggests that the MTD method of filler selection may yield better quality lineups, as well as improving witness identification accuracy.

While the MTD method was proposed to combat the issues with too much similarity,

problems with this method appear when witnesses provide vague descriptions or omit “default values” characteristics which appear obvious to both the witness and officer but need to be confirmed to create an unbiased lineup (Lindsay, Martin, & Webber, 1994). Examples of default values could include anything from race to gender depending on where a crime was committed and what that crime was. For instance, if the area where a crime was committed is predominantly one race and ethnicity, and the culprit is a member of that race and ethnicity, then the description of race may be omitted as the witness believes it is a given. Another example of this could be something like facial hair. If the culprit is clean-shaven, then witnesses may not think to mention this physical characteristic, or lack thereof, simply because it may seem obvious to them. Therefore, if vague descriptions are elicited and some of these default values are missing then lineup constructors may not have enough information to create a fair lineup which could cause the suspect to stand out. Therefore, the success of this method is at least partly contingent on eliciting high-quality witness descriptions.

Another problem with this method occurs when there are multiple witnesses to a crime and each provides slightly differing descriptions of the culprit. In this situation, officers would either need to construct a new lineup for each witness or create a modal description. Modal descriptions are created by selecting only the most frequently reported characteristics from all witness descriptions to include in the final description of the culprit (Malpass, Tredoux, McQuiston-Surrett, 2007). While it may seem redundant to distinguish between these two methods as the suspect should match the witness description, there are instances where a police suspect is either suspected due to other evidence or sometimes the police suspect is not the culprit. In these instances, the fillers selected for a lineup may not actually match the culprit description (Luus & Wells, 1991).

Perhaps because of the difficulty in sourcing sufficiently detailed descriptions of perpetrators, many studies use some combination of the description and the suspect's appearance to create lineups. The next section focuses on how the MTD has been implemented in research studies.

Implementing the Match-to-Description Method

There is much variation in the ways in which the MTD method has been implemented since it was first studied. Wells and colleagues (1993) were the first to propose the use of matching fillers to the culprit description provided by eyewitnesses. This was done by obtaining perpetrator descriptions from each witness and then creating lineups by selecting fillers based on how closely they matched those descriptions. Additionally, as a part of this proposed MTD method, Wells and colleagues (1993) selected new fillers and created a new, unique lineup to match each witness description rather than combining all witness descriptions to create one modal description. This variation of the method is what we consider to be a “pure” implementation of the MTD method of selecting fillers, as the fillers chosen for the final lineup are chosen *solely* based on how closely they match the culprit description. However, their proposed method differs greatly from how this method is often used in both research studies and in practice within the criminal justice system.

In many implementations of the MTD method, fillers are first selected from a large pool of options through use of the culprit description, and then a second round of filtering takes place. The fillers who make it through that final round of filtering are then used in the final lineup. Often, this filtering is accomplished using similarity ratings (Carlson, Gronlund, & Clark, 2008). In these studies, researchers use similarity ratings to compare the similarity of the fillers to the suspect and select those fillers with the highest similarity to the suspect to use in the lineup.

Other times similarity ratings are used to compare the similarity of the narrowed pool of fillers to the culprit description, as well as similarity to the other fillers, in order to make the final lineup selections (Haw & Fisher, 2004). We refer to these implementations, or variations, of the MTD method as “hybrid” implementations of the MTD method, as there is a subsequent filtering of fillers to make final filler selections to use in the lineup. However, these subsequent filtering approaches may lead to a bias when selecting the fillers for the suspect lineup. Specifically, subsequent filtering by appearance may undermine the utility of the MTD method by reintroducing the problem of too-high similarity.

Blinding of Lineup Constructors

Even if lineup constructors do not explicitly narrow the description-matched pool based on similarity to the suspect, it is still possible that knowledge of the suspect’s appearance may affect their filler selections. Top-down information processing explains how prior knowledge and experience can influence perception and decision making (Firestone & Scholl, 2014; Kinchla & Wolfe, 1979). This type of processing is conceptually driven, meaning it makes use of experience and existing knowledge to help individuals process and understand the world around them. One example of top-down processing is being able to see different shapes or figures out of ambiguous figures, such as the well-known duck/rabbit or younger/older woman figures (Nicholls, Churches, & Loetscher, 2018; Struber & Stadler, 1999). If you are told the figure is a duck, you will likely see the duck first based off that prior knowledge before further evaluation of the figure turns up the rabbit. Additionally, factors such as age, emotion, or mental state could influence top-down processing of incoming information (Firestone & Scholl, 2014; Proffitt, 2006). For example, Nicholls and colleagues (2018) found that when presented with the younger/older woman ambiguous figure, participant age influenced how young participants

perceived the face to be, suggesting how experience and perspective can influence decisions and how we process information. Thus, showing how experience, prior knowledge, or perspective can impact the views and decisions of an individual.

Knowing how top-down processing can influence decisions and processing of information is important to consider in various aspects of lineup procedure, including lineup administration and lineup construction. Much research has been done on blinding lineup administrators to the suspect's identity, as it has been discovered that when lineup administrators have knowledge of the suspect, they can unintentionally influence a witness's decision through verbal or nonverbal cues (Wells et al., 1998; Wells et al., 2020). This procedure of blinding the lineup administrator to the suspect is referred to as a double-blind lineup procedure (Kovera & Evelo, 2017). Research has shown that lineup administrators who know the identity of the suspect may allow their expectations for the eyewitness identification to bias the witness's identification decision. For example, research by Rhead et al. (2015) found that witnesses were more likely to make an identification from a lineup when the lineup administrator used a mix of verbal and nonverbal steering tactics to encourage witnesses to identify the suspect compared to witnesses who were administered a double-blind lineup. This finding is true for nonverbal cues as well, as demonstrated by research conducted by Haw and Fisher (2004) which found that witnesses were more likely to identify the suspect when the lineup administrator, who knew the identity of the suspect, was standing behind them rather than in front of or beside them. Further research supports this conclusion as witnesses who are administered single-blind lineups are more likely to make an identification than witnesses who are administered a double-blind lineup (Kovera & Evelo, 2017).

Additionally, decades of research have shown that lineup administrators who have

knowledge of the suspect may influence a witness's confidence in their identification decision, either intentionally or unintentionally, through post identification feedback (Wells et al., 2020). Research by Charman and Quiroz (2016) found that blinded and nonblinded lineup administrators behave differently. For example, nonblinded lineup constructors were more likely to exhibit nonverbal cues, such as smiling during or after an identification, than blinded administrators. This difference in behaviour led to witnesses making more false identifications and led them to being more confident in those false identifications. This finding demonstrates that blinding administrators plays an important role in the reduction of false identifications.

However, while much research has been done on blinding lineup administrators and the importance of administering double-blinded lineups, the impact of blinding lineup *constructors* to the suspect's appearance has not been systematically researched. One would expect the same reasoning that applies to blinding lineup administrators would apply to lineup constructors as once you know something, it is impossible to unknow it. If a lineup constructor knows who the suspect is and what they look like, any filler selections made using the MTD method may be biased by this knowledge. Crucially, this may occur even if they are attempting to use only the description to make their filler selections, as they will be reading and interpreting the description through the lens of this prior knowledge. Blinding a lineup constructor could prevent bias in the filler selection process, which could in turn lead to fairer lineups and reduce the number of eyewitness misidentifications. Thus, it may be prudent to blind lineup constructors as well as lineup administrators to the suspect in order to increase fairness and improve procedures related to eyewitness identifications.

Aims and Purpose

The purpose of this review is to better understand the MTD method of lineup

construction and if there are ways to improve this method, such as blinding the lineup constructor. During a review of the literature, it became apparent that there is currently no research on blinding the lineup constructor. This existing gap in the research begs the question, does blinding the constructor have any influence on lineup quality and fairness? This led us to the questions we plan to answer within this systematic review and meta-analysis: 1. How has the MTD method been operationalised in previous research? 2. Are lineup constructors in studies that have used an MTD method typically blinded or unblinded? 3. In studies that have used MTD methods with unblinded lineup constructors, how have the researchers attempted to prevent knowledge of the suspect's appearance biasing the filler selection process? 4. Have any previous studies manipulated blinding of the lineup constructor when using the MTD method? Through a meta-analysis, we will also answer the following question: is the MTD method associated with better lineup outcomes than the MTS method?

Answering these questions will provide a better understanding of how the MTD method has been operationalised throughout the years and how it compares to the more commonly used MTS method of filler selection. This knowledge could have profound implications on lineup policy and procedures if a fix as simple as blinding the lineup constructor to the suspect's appearance can improve lineup quality and fairness.

Method

Search Strategy

A pre-registered systematic review of the existing literature was conducted to investigate the influence of line-up construction methods on eyewitness identification reliability, and to investigate variability in how these construction methods have been implemented. Databases searched for the purposes of this review were ProQuest, SCOPUS, JSTOR, and Google Scholar.

The search was conducted between 18th October 2022 and 1st November 2022. The following Boolean search strings were used: “*suspect line-up construction*”; “*suspect line-up AND filler selection*”; “*suspect line-up AND filler similarity*”; “*propitious heterogeneity AND suspect line-up*”; “*match-to-description AND line-up*”; and “*match-to-suspect AND line-up*”.

In order to minimise the risk of publication bias, we contacted authors for unpublished work regarding the MTD lineup construction method. We contacted authors who had previously published research on filler selection and who have published lineup studies that have used the MTD method to create lineup stimuli. Authors were identified by looking at the correspondence information for the first authors on the articles that were included for data extraction. Additional author information was located for the second and third authors on these articles if applicable. We contacted authors via email and requested unpublished studies in which a pure implementation of the MTD method was used where the final selection of fillers was solely based on the description of the culprit. The unpublished studies also had to meet our inclusion criteria for the purpose of this review.

Inclusion and Exclusion Criteria: Systematic Review

The inclusion criteria for the initial title and abstract screening were: i) the MTD method of line-up construction was used to create some, or all, of the line-ups used in the study; ii) there was a known perpetrator; this includes staged/mock crime studies, and other studies that have included a known target; and iii) the study was written in English.

Screening Strategy

To ensure interrater reliability, CE and RH independently screened the titles and abstracts of the first 25 articles. Both CE and RH then compared decisions about articles to screen out and articles to escalate to full text review. Whenever disagreements about any of the articles

occurred, CE and RH discussed these disagreements, clarified the screening criteria (where necessary) and then screened the titles and abstracts of the next 25 articles independently. This was done for four separate batches until 100% agreement was reached, with Cohen's kappa being calculated to measure inter-rater reliability. A moderate level of agreement was found between the two raters, $\kappa = .54$, $p = < .001$. Once 100% agreement was reached, CE proceeded to screen the remaining titles and abstracts independently.

During full text review, the following exclusion criteria were applied: i) no MTD line-ups were used in the creation of line-ups used in the study; ii) there was insufficient detail provided about the line-up construction to determine whether the MTD method was used; iii) the study used stimuli from another published paper which was already captured; iv) the studies were published elsewhere, meaning the studies conducted a secondary analysis of previously published datasets; and v) there was no ground truth, or there was no known perpetrator. The same process described above was used to ensure interrater reliability, with CE and RH independently screening batches of 25 articles until all articles were screened as 100% agreement was not reached prior to completing the full text review. Cohen's kappa was calculated to measure inter-rater reliability and found a moderate level of agreement between the two raters, $\kappa = .48$, $p = < .001$.

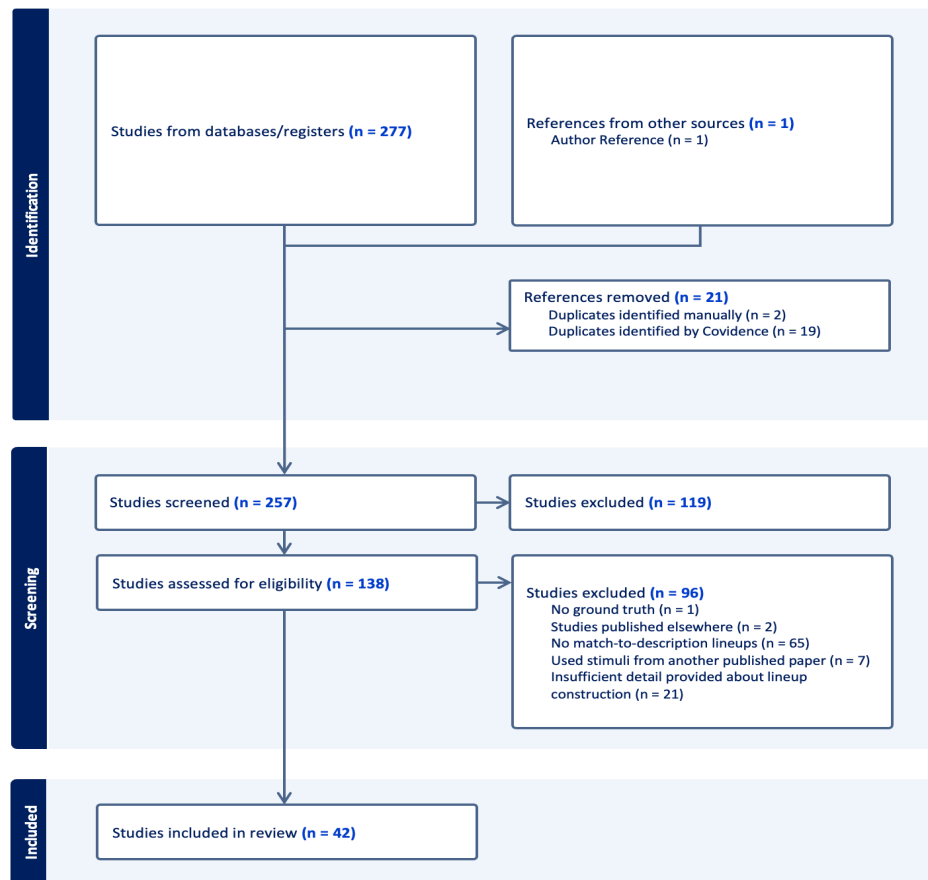
Data Extraction

For the data extraction stage, the following coding criteria were measured: a) if the lineup construction method was manipulated, or held constant; b) the number of lineup constructors used; c) whether the lineup constructor was blinded or not in the MTD condition; d) if the lineup constructor in the MTD condition was unblinded, whether any steps were taken to minimise potential bias resulting from the constructor having knowledge of the suspect's appearance; e) if

there was any subsequent selection or filtering of fillers after potential fillers were identified; f) the databases and sources used for filler selection; g) methods used to elicit descriptions; h) the number of witnesses that provided descriptions and how those descriptions were aggregated. Both CE and RH independently completed the data extraction template created in Covidence for the first five articles. Any discrepancies were discussed, the coding scheme was refined, and then data was extracted from another batch of five. Once in complete agreement, CE proceeded to complete the data extraction for the remaining articles. As seen in *Figure 1*, we began with 278 articles for review, with 42 articles meeting the necessary criteria and being escalated for data extraction.

Figure 1

PRISMA Flow-Chart



Systematic Review Results

Match-to-Description Implementations

Our first research question was concerned with ways in which the MTD method had been operationalised in previous research. The answer to this question is that the MTD method of lineup construction has been operationalised in several ways. From the original implementation of the method in which individual lineups were created for each witness description of the culprit and fillers were selected *solely* off those culprit descriptions, to newer implementations such as the iterative MTD method, it appears this method has been used in various ways throughout the years (Wells et al., 1993; Turtle et al., 2003). It is much more common to see hybrid methods used in the literature, where the description is used to create a pool of description-matched fillers but then there is a subsequent step (or steps) in which selections are based on appearance. For example, when officers or researchers enter a culprit description into a large database to narrow down the suspect pool and then use similarity-to-suspect ratings to create the lineups (Lucas, Brewer, & Palmer, 2021; Haw & Fisher, 2004). Another hybrid approach example is the *iterative MTD* method, where a culprit description is entered into a large face database and then the individual most similar to the suspect is selected as the first filler, the suspect's photo is then put to the side and the next filler is selected based on the first filler's appearance, this process continues until enough fillers have been selected for the lineup (Turtle, Lindsay, & Wells, 2003).

Of the 42 articles which described their lineup construction method as "match-to-description", 26 used a method in which the final selection of fillers was based *only* on the description of the perpetrator, please refer to *Table 1*. The remaining articles used what may be considered a hybrid approach or combined version of the MTD and MTS methods where the

MTD method was used for the initial narrowing of fillers, but the subsequent filtering and selection of fillers was through another technique. These subsequent filtering methods may include similarity to suspect ratings, mock witness tasks, and the previously mentioned iterative MTD method.

Table 1

Match-to-Description Implementations

Study ID	MTD method used?	Lineup Constructor Blinded?	Subsequent Filtering?	Subsequent Filtering Method Used
Kempen 2009	Yes (Pure)	Unclear	Unclear	Unclear
Colloff, Wade, & Strange 2016	Yes (Pure)	Unclear	Unclear	Unclear
Hope, Lewinski, Dixon, Blocksidge, & Gabbert 2012	Yes (Pure)	Unclear	Unclear	Unclear
Charman & Cahill 2012	Yes (Pure)	Unclear	Unclear	Unclear
Jones 2021	Yes (Pure)	Unclear	Unclear	Unclear
Mitchell 2005	Yes (Pure)	Unclear	Unclear	Unclear
Wells, Rydell, & Seelau 1993	Yes (Pure)	Yes	No	N/A
Smith et al. 2014	Yes (Pure)	Yes	No	N/A
Tunnicliff & Clark 2000	Yes (Pure)	Yes	No	N/A
Wilcock & Bull 2010	Yes (Pure)	Yes	No	N/A
Tredoux et al. 2007	Yes (Pure)	Unclear	No	N/A
Darling, Valentine, & Memon 2008	Yes (Pure)	Unclear	Unclear	Unclear
Wood 2017	Yes (Pure)	Yes	No	N/A
MacLin, Zimmerman, & Malpass 2005	Yes (Pure)	Unclear	Unclear	Unclear
Brewer, Weber, & Guerin 2020	Yes (Pure)	Unclear	Unclear	Unclear
Stinson et al. 1997	Yes (Pure)	Unclear	No	N/A
Brewer et al. 2012	Yes (Pure)	Yes	No	N/A
Mansour et al 2009	Yes (Pure)	Unclear	No	N/A
Rubínová et al 2021	Yes (Pure)	Unclear	Unclear	Unclear
Wittwer et al 2022	Yes (Pure)	Yes	No	N/A
Weber & Varga 2012	Yes (Pure)	Unclear	No	N/A
Holdstock et al 2022	Yes (Pure)	Unclear	Unclear	Unclear
Wilson et al 2019	Yes (Pure)	Unclear	Unclear	Unclear

Guerin et al. 2018	Yes (Pure)	Unclear	Unclear	Unclear
Russ et al 2018	Yes (Pure)	Unclear	Unclear	Unclear
Juslin, Olsson, & Winman 1996	Yes (Pure)	Unclear	No	N/A
Brewer et al 2022	Yes (Hybrid)	Unclear	Yes	
Brewer & Wells 2006	Yes (Hybrid)	Unclear	Yes	Used a combination of face similarity ratings and the distribution of lineup foil choices recorded in a previous study
Horry, Brewer, & Weber 2016	Yes (Hybrid)	Unclear	Yes	Similarity rating procedure
Gronlund et al 2009	Yes (Hybrid)	Unclear	Yes	Similarity rating procedure
Carlson, Gronlund, & Clark 2008	Yes (Hybrid)	No	Yes	Similarity ratings used
Carlson et al 2019	Yes (Hybrid)	Unclear	Yes	
Lindsay et al 1994	Yes (Hybrid)	No	Yes	
Lucas, Brewer, & Palmer 2021	Yes (Hybrid)	Unclear	Yes	Similarity rating procedure
Dysart, Lawson, & Rainey 2012	Yes (Pure)	Unclear	Unclear	Unclear
Parliament & Yarmey 2002	Yes (Hybrid)	Unclear	Yes	
Sučić, Tokić, & Ivešić 2015	Yes (Hybrid)	Unclear	Yes	A group of mock witnesses (N= 16) estimated the similarity between each pair of photographs by using a 7-point scale, and another group (N = 27) estimated the similarity of these photographs to the modal description of the target person by using a 7-point scale.
Haw & Fisher 2004	Yes (Hybrid)	Unclear	Yes	Similarity rating procedure
Goodsell, Neuschatz, & Gronlund 2009	Yes (Hybrid)	Unclear	Yes	Similarity rating procedure
Haw 2005	Yes (Hybrid)	Unclear	Yes	
Pike et al 2019	Yes (Hybrid)	Unclear	Yes	

Blinding of the Lineup Constructor

While there have been several ways in which the MTD method has been operationalized, the lack in transparency in reporting makes it difficult to determine exactly how this method is

implemented in many cases. One potentially important detail, which is rarely reported, is whether the lineup constructor is blinded to the suspect's appearance. In 7 of the 42 studies, 16.67%, we were able to determine the constructor was blinded, while 2 of the 42, 4.76%, reported lineup constructors were not blinded to the suspect's appearance during the filler selection process. Though 9 of the 42 articles escalated for data extraction have reported information regarding blinding of lineup constructors, the majority of studies do not report this information. This lack of transparency in reporting was the case for 33 of the 42 studies in which data was extracted, that is approximately 79% of articles not reporting important information regarding the filler selection process and lineup construction. Therefore, while it appears some studies have operationalised what we would consider to be a "pure" implementation of the MTD method, it is unclear whether this implementation of the method improves lineup fairness and eyewitness reliability without more transparent reporting and more research done on this aspect of lineup construction and filler selection.

Reducing Information Bias

In instances where the lineup constructor was not blinded to the suspect's appearance, the most common method of selecting fillers, after an initial description-matched pool was created, was to select fillers using similarity-to-suspect ratings (Carlson et al., 2008; Goodsell et al., 2009; Lucas et al., 2021). However, there was one instance in which the final selection of fillers conducted using similarity-to-description ratings (Sucic et al., 2015). In this study, two groups of mock witnesses provided similarity ratings of the matched-to-description pool of potential fillers. The first group rated each of the photos for similarity among the description matched suspect pool, while the second group rated these photos from the description match suspect pool for similarity to the modal description. For target absent lineups, the six photographs that were

selected were those that ranked highest for photograph similarity and match to modal description. The photograph ranked highest by both groups became the innocent suspect and was replaced in the lineup by the target for target present lineups.

While this method of subsequent filler selection through similarity to description ratings seems beneficial in reducing bias from unblinded lineup constructors, most studies which had a subsequent filtering of fillers did so by using similarity-to-suspect ratings. For example, in some studies, after the initial narrowing of potential fillers using the MTD method, independent observers were brought in to rate the faces based on similarity to the target and those rated most similar were used in the lineup (Gronlund et al., 2009; Brewer & Wells, 2006). Another similar method commonly used by researchers is to create a narrowed-down, description-matched pool and from there selecting fillers based off perceived similarity to the suspect (Darling, Valentine, & Memon, 2008). Studies using these methods of filler selection bring to question whether these processes of filler selection should be considered a hybrid approach to the MTD method or if they fall into the MTS method of lineup construction. It also brings to light whether these methods truly reduce the bias on filler selection by not blinding the lineup constructor to the suspect's appearance.

While it is apparent there is variability in the ways in which the MTD method has been implemented, it is hard to know how much variability exists due to the lack of transparency in reporting (*Table 1*). Lack of reporting also makes it difficult to correctly categorize whether a study has used the MTD method or some combination of description and suspect-matched methods. In addition, this lack of transparency makes it difficult to determine how well the practices used in controlled, lab settings under the guise of being a description-matched procedure align with how law enforcement officials implement this method.

Risk of Bias

Risk of bias was assessed by adapting the *Cochrane Risk of Bias tool (RoB 2)* to be suitable for assessing social and behavioural science research (*Table 2*) (Sterne et al., 2019). Each Risk of Bias item was scored as low risk (0), unclear (1), or high risk (2). Individual item scores were then summed to create total Risk of Bias scores ranging from 0 to 14. The Risk of Bias scores for each study included in the systematic review can be found in *Table 3*. Due to a lack of transparency in reporting, it was difficult to determine risk criteria for certain risk items such as Blinding of Analysts. As seen in *Table 3*, the risk was deemed to be low for most of the risk items with the exception being for the ‘Other Sources of Bias’ risk item. All but one study were deemed to be high risk for this item as none of these studies mentioned being pre-registered and following pre-specified research plans.

Table 2

Risk of Bias Criteria. Adapted from Horry et al. (2021).

Bias Domain	Risk Item	Criteria
Selection Bias	Random allocation	Low risk: Participants were randomly allocated to experimental conditions High risk: Participants were not randomly allocated to experimental conditions Unclear: Reporting is not clear, unable to determine if participants randomly allocated
	Allocation concealment	Low risk: The experimenter did not know which experimental condition the participant was in prior to any interactions with the participant High risk: The experimenter knew which experimental condition the participant was in prior to any interactions with the participant Unclear: Reporting not clear if experimenter had previous knowledge of what condition participant was in
Reporting Bias	Selective reporting	Low risk: (Where a protocol exists), the researchers reported all the dependent measures/analyses included in their analysis plan; all outcome measures listed in the Methods sections are reported in the Results section; non-significant results are reported fully High risk: (Where a protocol exists), the researchers did not report all of the dependent measures/analyses included in their analysis plan; outcome measures were listed in the Methods section, but analyses were not reported; non-significant results were either not reported fully (e.g., n.s. or $p > .05$) or were not reported at all. Unclear

Performance Bias	Blinding participants	Low risk: Participants were blinded to the lineup construction method that was used High risk: Participants were not blinded to the lineup construction method that was used Unclear: Not clear if participants were blinded to construction method Not Applicable - only one lineup construction method was used
Detection Bias	Blinding analysts	Low risk: Data analyst is blinded to experimental conditions High risk: Data analyst is not blinded to experimental conditions Unclear: Not clear if analyst was blinded
Attrition Bias	Incomplete outcome/ withdrawals/missing data	Low risk: All participants who were recruited were included in the analyses; The number of participants who withdrew from the study is clearly reported; The number of participants in each experimental condition is reported; Exclusions are clearly reported and justified High risk: The number of participants recruited does not match the number of participants included in the analysis; Participant withdrawals are not clearly reported; The number of participants in each experimental condition is not clearly reported; Exclusions are not clearly reported and/or are not adequately justified Unclear
Other Bias	Other sources of bias	Low risk: Research questions and analysis plans were pre-specified using a timestamped pre-registration High risk: Research questions and analysis plans were not pre-specified using a timestamped pre-registration Unclear

Table 3

Risk of Bias Scores for Each Study Included in Systematic Review

Study ID	Random allocation	Allocation concealment	Blinding participants	Blinding of analysts	Incomplete outcome data	Selective reporting	Other sources of bias	Total Risk of Bias
Kempen 2009	0	1	0	1	0	0	2	4
Colloff et al. 2016	0	1	0	1	0	2	2	6
Hope 2012	0	1	0	1	0	0	2	4
Charman 2012	0	1	0	1	0	0	2	4
Jones 2021	0	1	1	1	0	0	2	5
Mitchell 2005	0	1	0	1	0	2	2	6
Wells 1993	0	2	0	1	0	2	2	7
Brewer 2006	0	1	0	1	0	2	2	6
Horry 2016	0	1	0	1	0	2	2	6
Palmer 2012	0	1	0	1	0	0	2	4
Tunnicliff 2000	0	0	0	1	0	2	2	5
Darling 2008	1	1	1	1	0	2	2	8

Gronlund 2009	0	0	0	1	0	2	2	5
MacLin 2005	0	1	0	1	0	0	2	4
Brewer 2020	1	1	0	1	0	2	2	7
Wilcock 2010	0	1	0	1	0	0	2	4
Tredoux 2007	0	0	0	1	0	0	2	3
Wood 2017	0	0	0	1	0	0	2	3
Rubinova 2021	0	0	0	1	0	0	2	3
Carlson 2008	0	0	0	1	0	0	2	3
Carlson 2019	0	0	0	1	0	0	2	3
Holdstock 2022	0	0	0	1	0	0	0	1
Wilson 2019	0	0	0	1	0	0	2	3
Guerin 2018	0	0	0	1	0	0	2	3
Russ 2018	1	2	0	1	0	0	2	6
Stinson 1997	0	0	0	1	0	0	2	3
Brewer 2012	0	1	0	1	0	0	2	4
Lindsay 1994	2	1	0	1	1	2	2	9
Lucas 2021	0	0	0	1	0	0	2	3
Parliament 2002	0	0	0	1	0	2	2	5
Mansour 2009	0	0	0	1	0	2	2	5
Sucic 2015	0	2	0	1	1	0	2	6
Wittwer 2022	0	0	0	1	0	0	2	3
Brewer 2022	0	0	0	1	2	0	2	5
Dysart 2012	0	1	0	1	0	0	2	4
Haw 2004	0	1	0	1	0	2	2	6
Goodshell 2009	0	1	0	1	2	2	2	8
Weber 2012	0	1	0	1	1	0	2	5
Juslin 1996	1	1	0	1	0	0	2	5
Haw 2005	0	1	0	1	0	0	2	4
Carlson 2019	0	0	0	1	0	0	2	3
Pike 2019	0	1	0	1	0	2	2	6

Note: RoB codes: 0 = Low Risk, 1 = Unclear, 2 = High Risk. Higher RoB scores indicate a

higher risk of bias.

Discussion

The purpose of this systematic review was to 1) better understand how the MTD method of lineup construction has been implemented and 2) to determine if there are ways to improve this method, such as blinding the lineup constructor. The results of this systematic review

suggest that the MTD method has been operationalised in ways which vary greatly from its original implementation. From its original implementation in which fillers were selected *solely* on the perpetrator description, to newer implementations such as the iterative MTD method, it appears the MTD method of lineup construction has been used in various ways throughout the years (Wells et al., 1993; Turtle et al., 2003). The findings of this review revealed the most common implementation of the MTD method is a hybrid approach in which the perpetrator description is used to create a pool of description-matched fillers but then there is a subsequent step (or steps) in which selections are based on the suspect's appearance.

In addition to understanding how the MTD method of lineup construction has been operationalised, we also wanted to understand how blinding MTD lineup constructors to the suspect's appearance may influence the effectiveness of the method. However, a lack of transparency in reporting made it difficult to determine whether researchers had blinded their MTD lineup constructors to the suspect's appearance. Only 9 of the 42 articles escalated for data extraction reported information regarding blinding of lineup constructors. This means approximately 79% of articles included in this systematic review did not report important information regarding the filler selection process and construction of suspect lineups. With that said, it makes it difficult to determine exactly how this method has been operationalised by researchers and whether blinding MTD lineup constructors would improve lineup fairness and eyewitness reliability.

The review also revealed there are currently no studies in which the blinding of lineup constructor using the MTD method has been manipulated. As it appears the MTD method is often implemented using a hybrid approach, it brings to question whether the frequently reported non-significant differences between MTD and MTS lineups could be due to this lack of blinding

lineup constructors. With that said, the research regarding filler selection method remains conflicted as to which method yields better lineup outcomes. Thus, as eyewitness misidentifications continue to play a role in around 25% of all wrongful convictions in both the United States and United Kingdom, it is important to find ways to improve the ways in which suspect lineups are constructed and administrated (National Registry of Exonerations, 2025; Miscarriages of Justice Registry, 2025). Therefore, these gaps in the research need investigating to determine which construction method is most appropriate for creating fair suspect lineups. In order to provide additional insight into how the MTD method of lineup construction performs compared to the MTS method, a meta-analysis was conducted.

Meta-analysis

While the purpose of the systematic review was to understand how the MTD method of lineup construction has been implemented over the years, the purpose of this meta-analysis is to examine how the MTD method of lineup construction method compares to the MTS method of lineup construction in terms of lineup outcomes. This meta-analysis aimed to answer the following question: does the MTD method improve lineup outcomes? As the existing literature remains conflicted as to which lineup construction method yields fairer, better quality lineups, it is apparent additional research is needed. This meta-analysis provided needed insight into which method leads to better eyewitness identification decisions. It may also give us insight into whether lineups created using the “pure” MTD method perform better than those created using a “hybrid” MTD method.

Meta-analysis Method

Inclusion Criteria: Meta-Analysis

For an article to be included in the meta-analysis, the following criteria was used: i) the

study had to experimentally compare MTD lineups and MTS lineups; ii) the study had to be a lineup study in which participants selected an individual from a lineup where there was a known target or innocent suspect (i.e., not a lineup fairness study); iii) the article had fully reported results, including how many participant-witnesses were in each condition. This resulted in 7 of the 42 articles included in the systematic review being suitable for inclusion in a meta-analysis.

Meta-Analytic Procedure

Outcome Measures

Each of the 7 articles that met the inclusion criteria were reviewed and the data was extracted. Separate analyses were then conducted for each identification decision for target present and target absent lineups (i.e. suspect identifications, filler identifications, non-identifications/correct rejections). Only one study did not include a designated innocent suspect, as such this study was excluded for the target absent analyses. In addition to running analyses on identification decisions for target present and target absent lineups, we also calculated choosing rates and discriminability. Choosing rates were calculated for both MTD and MTS lineups for each article as odds ratios. Discriminability was calculated using the formula $d' = zH - zFA$ (MacMillan & Creelman, 1991; Mickes et al., 2014), where H refers to the hit rate (guilty suspect identifications) and FA refers to the false alarm rate (innocent suspect identifications). In order to compare discriminability across conditions, we used the Gourevitch and Galanter (1967) method to estimate the variance of a single d' score. This then allowed us to calculate the necessary inferential statistics to compare two group d' scores.

Moderator Analyses

A subgroup analysis was conducted through IBM SPSS Statistics (Version 30) using the restricted maximum likelihood (REML) estimation method to examine each of the five potential

moderator variables identified. The potential moderator variables identified include: i) the year the study was conducted, separated into pre- and post-2012. This cut-off was chosen because 2012 marked the beginning of the replication crisis. Consequently, studies conducted post-2012 tend to have increased sample sizes and greater transparency in reporting methodology; ii) MTD type, separated into what we classified as pure MTD implementation and hybrid MTD implementation; iii) subsequent filtering used to select the final fillers included in a lineup, separated into ‘unclear, yes, and no’; iv) participant type, either community volunteer or university student; v) and study type, the study conducted either online or in-person. These variables were chosen as potential moderators as at least 2 of the 7 studies included in this meta-analysis had differing responses from the others.

Outliers

Only one outlier was identified in this analysis. The Wells et al. (1993) study had a much larger effect size compared to the other studies included in this meta-analysis. Additionally, no other study had an effect size that fell within the bounds of the confidence interval for the outlier. All analyses were therefore conducted both with and without the outlier to ensure that this single study did not exert undue influence on the conclusions.

Publication Bias

In efforts to minimise publication bias, we reached out to authors who had previously published work on lineup construction methods for any unpublished work they may have regarding the topic. It is worth noting that by only contacting those individuals who had previously published research in the area, we may have missed other unpublished works relating to filler selection such as theses, dissertations, or conference presentations. These forms of “grey literature” (i.e. unpublished work or research that is produced and published by non-commercial

organisations) are often missing from systematic reviews and meta-analyses as there is no centralized database to locate this information and may contribute to publication bias (Bellefontaine & Lee, 2013). Therefore, while efforts were made to minimise publication bias, it is possible additional bias was introduced by not locating and accessing other forms of “grey literature”.

Publication bias analysis was conducted for each identification decision for both target present and target absent lineups. Publication bias was first assessed visually using funnel plots then using the trim-and-fill procedure. The trim-and-fill method uses statistical modeling to detect studies contributing to funnel plot asymmetry and then corrects this by imputing placeholders where studies appear to be missing (Duval & Tweedie, 2000; Juncu & Fitzgerald, 2021). These analyses suggested no publication bias for any identification decision except for filler identifications made from target present lineups. The funnel plots can be found in *Appendix A*.

Meta-analysis Results

Target Present Lineups

Our first set of analyses focused on lineup decisions from target-present lineups. We conducted separate analyses for correct identifications, filler identifications, and non-identifications. The results for correct identifications are shown in *Figure 2*. The weighted meta-analytic effect size was significantly greater than zero, $\ln OR = 0.53$ [0.09, 0.97], $p = .02$. In other words, the odds of a correct identification is approximately 1.70 [1.10, 2.61] times greater from lineups created using the MTD method than from lineup created using the MTS method. The between-study heterogeneity variance for correct suspect identifications was estimated at $\tau^2 = 0.22$, with an I^2 value of 73%, suggesting considerable heterogeneity. This is further supported

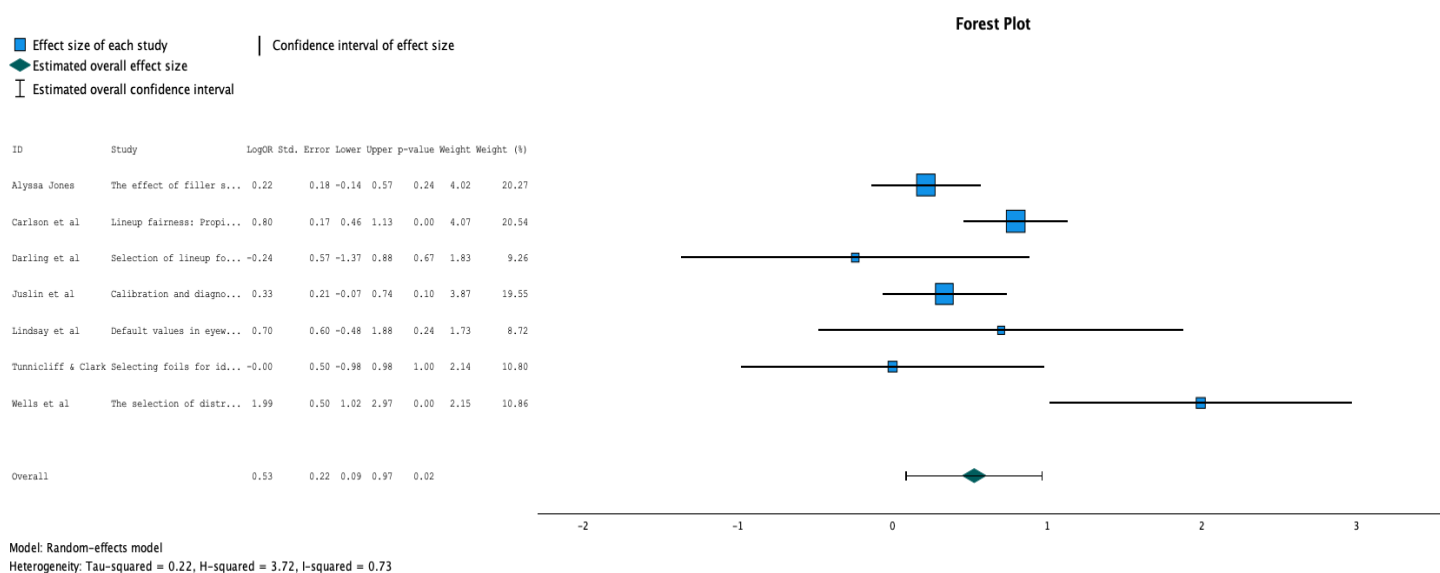
by $H^2 = 3.72$, as H^2 values over 1 indicate between-study heterogeneity.

One possible explanation for this considerable amount of heterogeneity is the study conducted by Wells and colleagues (1993) is an outlier in this analysis. Not only does it have a much larger effect size compared to the other studies included in this meta-analysis $\ln OR = 1.99$ [1.02, 2.97], $p < .001$, but additionally no other study has an effect size that falls within the bounds of the confidence interval for this study. With that said, it is important to note that nearly all of the point estimates are positive, but that for most studies, the confidence intervals overlap zero, unlike the Wells et al. (1993) outlier. This is especially true for the smaller studies, which have very wide confidence intervals.

Even once the Wells et al. (1993) outlier has been removed, these findings remain significant, $\ln OR = 0.39$ [0.09, 0.69], $p = .01$. Additionally, with this outlier removed we still have heterogeneity, $\tau^2 = 0.05$, an I^2 value of 44%, $H^2 = 1.77$. While it is slightly less than when it is included, the heterogeneity found demonstrated the need for a moderator analysis.

Figure 2

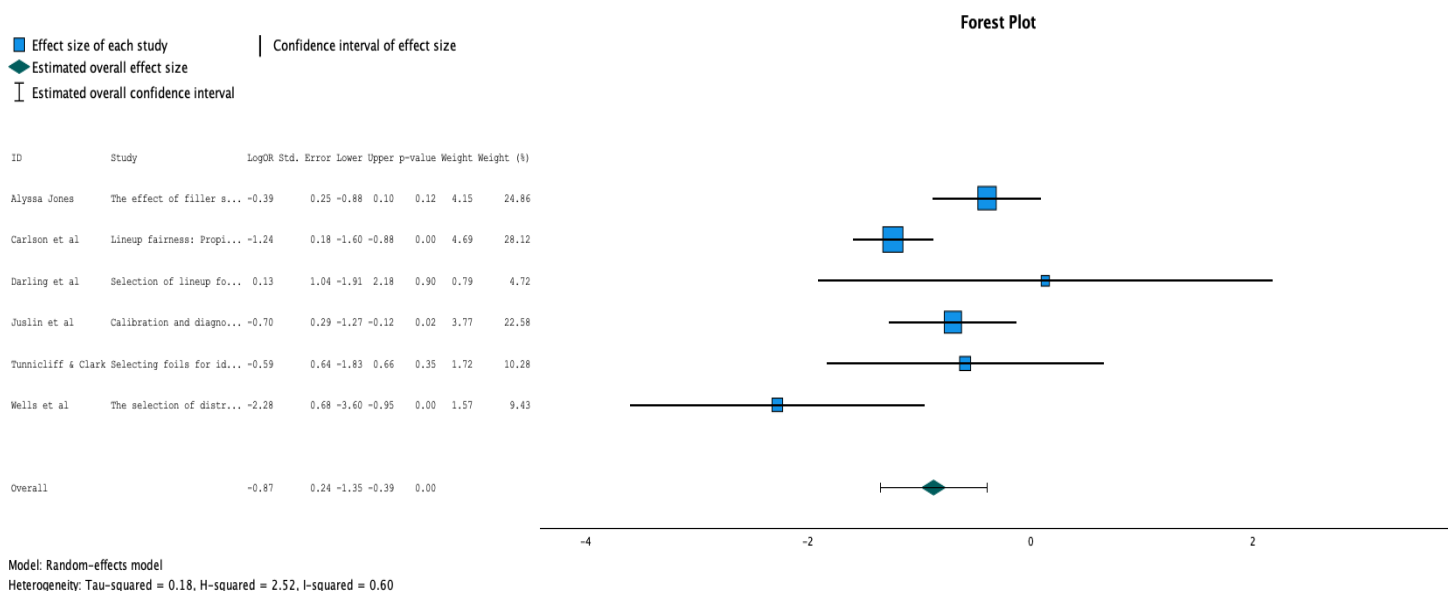
Forest Plot Depicting Correct Identifications for Target Present Lineups



The odds of a witness making an incorrect filler identification from a target-present lineup were lower when fillers were selected using the MTD method than when they were selected using the MTS method (*Figure 3*), $\ln OR = -0.87 [-1.35, -0.39]$, $p < .001$. In other words, the odds of making an incorrect filler identification are approximately 0.42 [0.26, 0.67] times greater in instances where lineups were created using the MTS method than those lineups created using the MTD method. The between-study heterogeneity variance for innocent filler identifications in target present conditions was estimated at $\tau^2 = 0.18$, with an I^2 value of 60%, suggesting considerable between-study heterogeneity. This is further supported by the H^2 value being greater than one, $H^2 = 2.52$.

As with above, one possible explanation for the amount of heterogeneity seen is the study conducted by Wells and colleagues (1993), which is an outlier in this analysis. Not only does it have a much larger effect size compared to the other studies included in this meta-analysis $\ln OR = -2.28 [-3.60, -0.95]$, $p < .001$, but only one other study has an effect size which falls within the bounds of the confidence interval for this study with the other studies having values much closer to zero.

However, the findings still hold true even with the Wells et al. (1993) outlier removed. There is still a significant main effect suggesting witnesses are more likely to select an innocent filler from a lineup made using the MTS method, $\ln OR = -0.74 [-1.19, -0.30]$, $p < .001$. Even with this outlier removed we still have heterogeneity, $\tau^2 = 0.12$, $I^2 = 0.55$, $H^2 = 2.21$, while it is slightly less than having it included, it demonstrated the need for a moderator analysis.

Figure 3*Forest Plot Depicting Innocent Filler Identifications for Target Present Lineups*

For non-identifications from target-present lineups, the meta-analytic effect size was not significantly different from zero $\ln OR = 0.14 [-0.09, 0.37]$, $p = .23$ (Figure 4). It is important to note that the only study resulting in fewer non-identifications when using the MTD method of lineup construction was the study conducted by Wells and colleagues (1993). The between-study heterogeneity variance for non-identifications in target present conditions was estimated at $\tau^2 = 0.00$, with an I^2 value of 0%, which suggests no heterogeneity. This is further supported by $H^2 = 1.00$, as H^2 values of 1 and below suggests no between-study heterogeneity. However, it is important to note that any heterogeneity that may have been seen from the Wells et al. (1993) outlier may not be reflected due to the wide confidence intervals causing much overlapping between studies.

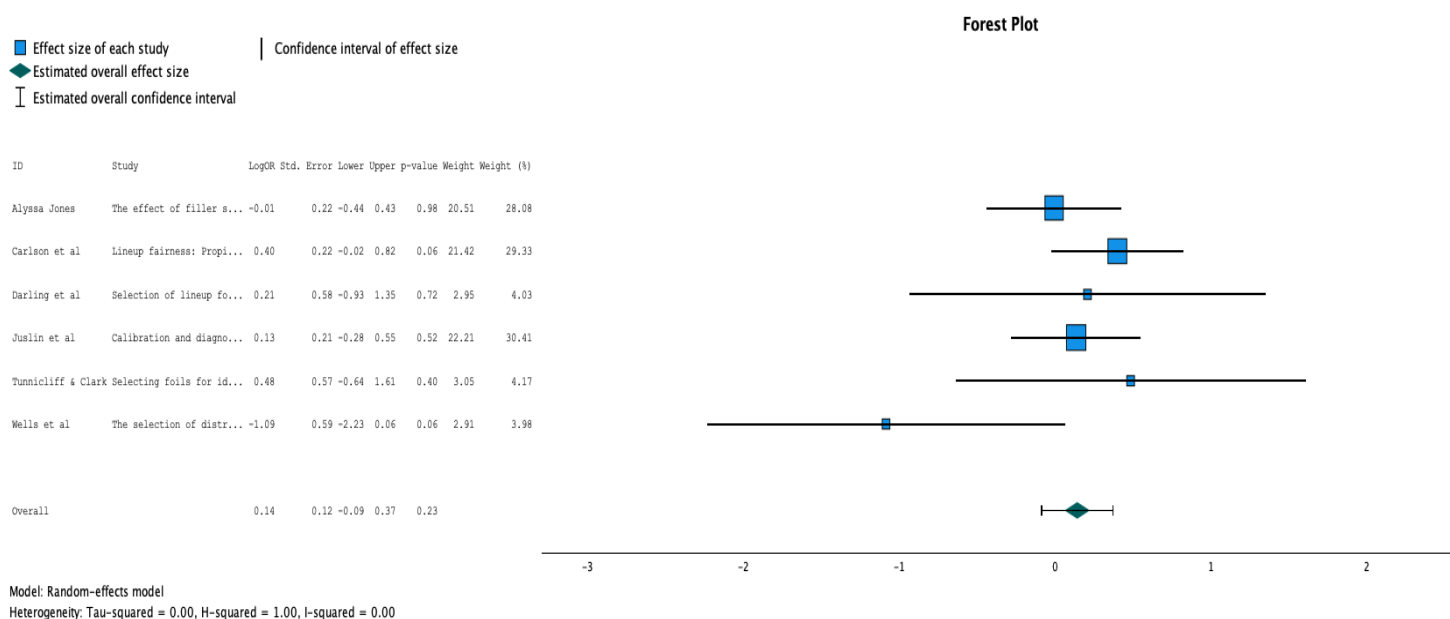
The findings for non-identifications from target-present lineups with the Wells et al. (1993) outlier removed are still not significant, $\ln OR = 0.19 [-0.04, 0.43]$, $p = .11$. Additionally, the between-study heterogeneity variance remains, $\tau^2 = 0.00$, with an I^2 value of 0%, which

suggests no heterogeneity. This is further supported by $H^2 = 1.00$, as H^2 values of 1 and below suggests no between-study heterogeneity. Even with this outlier removed, considerable overlap in confidence intervals remains.

The findings of this analysis suggest that lineups created using the MTD method, at least when the target is present, result in greater witness accuracy. While there does not appear to be a significant difference in non-identifications between the two lineup construction methods, we see an increase in correct identifications with description-matched lineups while also seeing a decrease in filler identifications. This suggests that perhaps, when the target is present, description-matched lineups improve a witness' ability to discriminate between a previously seen target and innocent fillers. With this said, further research is needed to confirm whether description-matched or suspect-matched lineups yield fairer lineups and increase witness reliability.

Figure 4

Forest Plot Depicting Incorrect Non-Identifications for Target Present Lineups



Target Absent Lineups

We also conducted a set of analyses focused on lineup decisions from target-absent lineups. We conducted separate analyses for innocent suspect identifications, filler identifications, and non-identifications. Six studies were included in the analyses of target absent lineup outcomes, as these each included a designated innocent suspect. For innocent suspect identifications, there was a significant effect of lineup construction method, $\ln OR = 0.34$ [0.06, 0.62], $p = .02$ (*Figure 5*). The odds of an innocent suspect identification was 1.4 [1.07, 1.85] times higher for lineups created using the MTD method than for those created using the MTS method. With that said, this finding should be considered with caution as it was influenced by one heavily weighted study which accounted for 83.49% of the weight, leaving the other five studies to account for the remaining 16.51% (*Figure 5*). This is due to the fact many studies included in the meta-analysis have used small sample sizes, which results in a large degree of uncertainty in effect sizes. Therefore, in order to make a more definitive response to the question regarding how lineup construction method impacts witness identifications, we need more research with high statistical power. With further research, we may be able to determine whether there is an effect of lineup construction method on innocent suspect identifications when the target is absent.

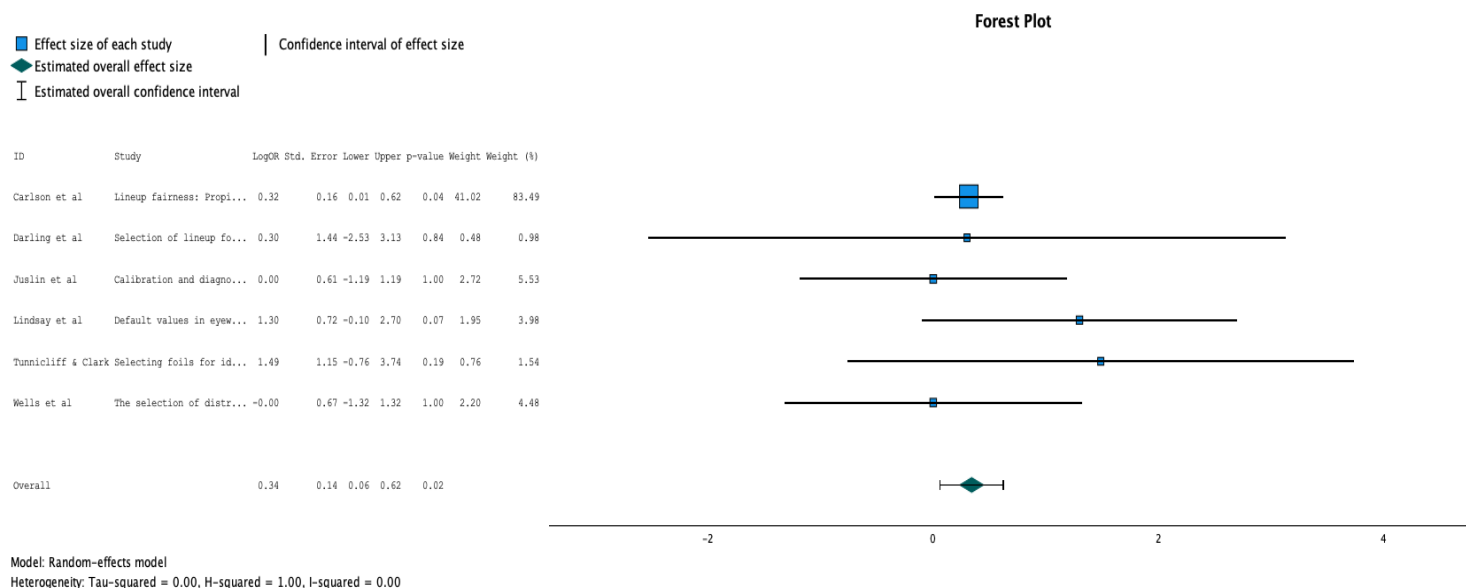
The between-study heterogeneity variance for innocent suspect identifications in target absent conditions was estimated at $\tau^2 = 0.00$, with an I^2 value of 0%, which suggests no heterogeneity. This is further supported by $H^2 = 1.00$, which as previously mentioned also suggests no between-study heterogeneity. While the I^2 statistic is commonly used to estimate the variance due to heterogeneity, it can be biased when there is small number of studies and should be interpreted with caution (von Hippel, 2015). This finding is due to the wide overlapping of the

confidence intervals which is a function of low sample size and a relatively small set of studies.

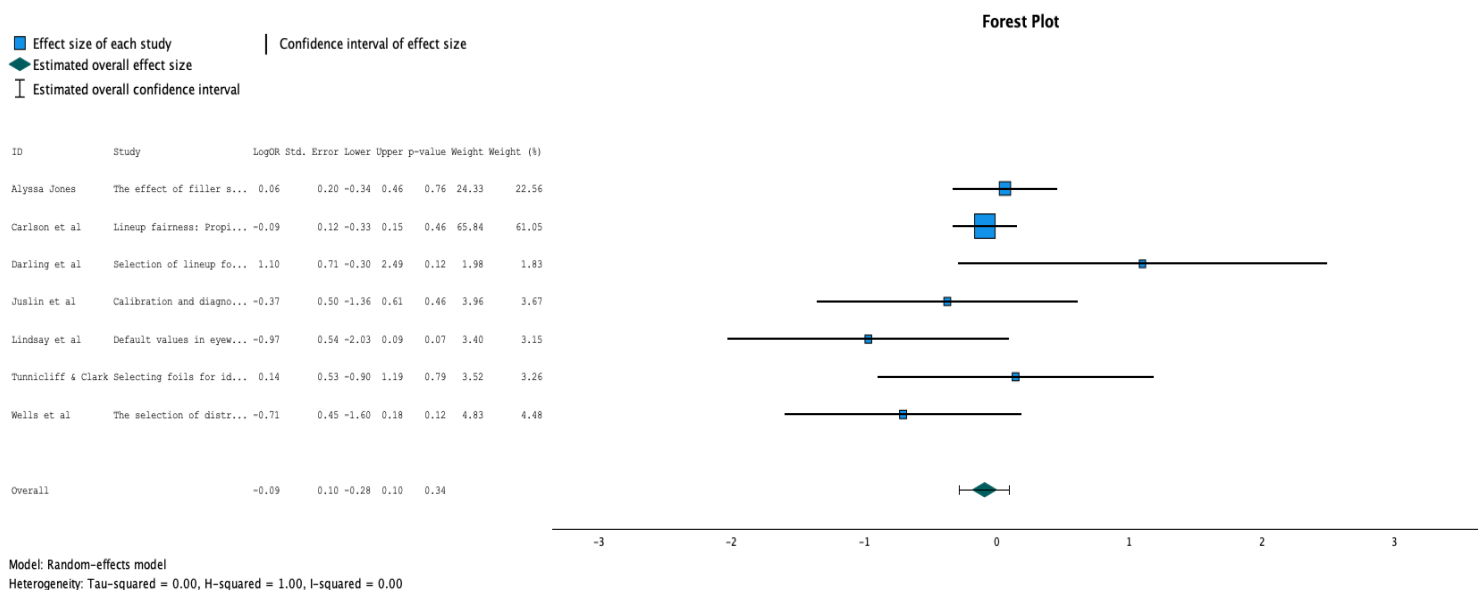
In analyses where there is a limited number of studies, looking for nonoverlapping confidence intervals is a more reliable indicator of between-study heterogeneity.

Figure 5

Forest Plot Depicting Innocent Suspect Identifications for Target Absent Lineups



There were non-significant effects for both innocent filler identifications and correct rejections. As seen in *Figure 6*, it is only slightly more likely (approximately 0.9 [0.75, 1.11] times greater) for an innocent filler to be identified when lineups are created using the MTS method than those using the MTD method, $\ln\text{OR} = -0.09 [-0.28, 0.10]$, $p = .34$. The between-study heterogeneity variance for innocent filler identifications in target absent conditions was estimated at $\tau^2 = 0.00$, with an I^2 value of 0% and $H^2 = 1.00$, both of which suggest no between-study heterogeneity.

Figure 6*Forest Plot Depicting Innocent Filler Identifications for Target Absent Lineups*

Additionally, there was a non-significant main effect for correct rejections, $\ln\text{OR} = -0.08$ $[-0.27, 0.10]$, $p = .38$, when the target was absent (*Figure 7*). In other words, the odds of making a correct rejection, or non-identification, is approximately 0.9 $[0.77, 1.10]$ times greater from lineups created using the MTS method than from lineup created using the MTD method. The between-study heterogeneity variance for non-identifications in target absent conditions was estimated at $\tau^2 = 0.00$, with and I^2 value of 0% and $H^2 = 1.00$, both values suggest no heterogeneity.

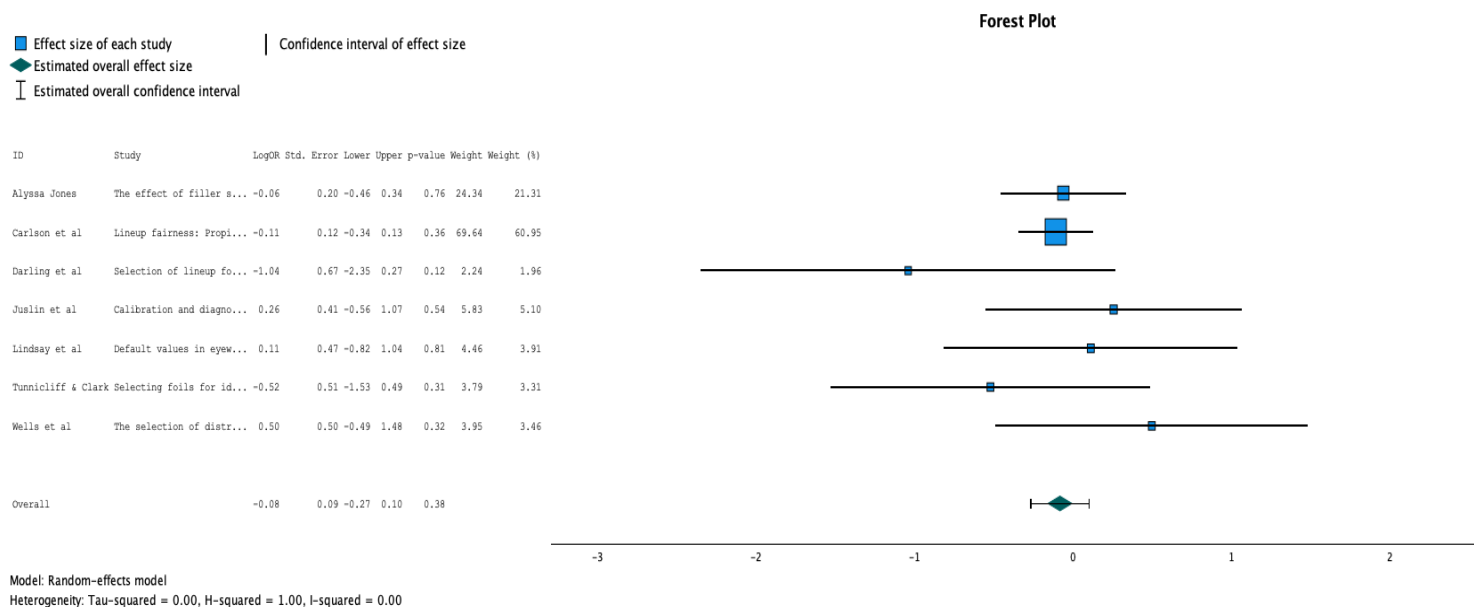
The results of this analysis suggest there is little difference in witness accuracy between description-matched and suspect-matched lineups when the target is absent, apart from innocent suspect identifications. While the analysis suggests innocent suspect identifications occur at a higher rate in description-matched compared to suspect-matched lineups, this finding was driven by a single, heavily weighted study. The studies included in this analysis simply do not have the

statistical power to determine the impact of lineup construction method on target absent lineups.

Therefore, in order to confirm whether description-matched lineups result in unfair lineups and reduced witness accuracy further research with adequate statistical power is required.

Figure 7

Forest Plot Depicting Correct Non-Identifications for Target Absent Lineups



Moderator Analyses

During the meta-analysis, significant between-study heterogeneity in effect size was found for correct suspect identifications and incorrect filler identifications when the target was present, which prompted the need for a moderator analysis. The I^2 value for correct suspect identifications was 73%, suggesting considerable heterogeneity between studies while the I^2 value for incorrect filler identifications was 60%, which also suggests considerable between-study heterogeneity. It is important to note that only seven studies met the inclusion criteria for the meta-analysis and should be considered when interpreting the findings of the moderator analysis. With so few studies, there is considerable overlap which is demonstrated by the

overlapping of the confidence intervals.¹

The subgroup analysis for correct suspect identifications in target present lineups found that when the Wells et al. (1993) outlier was included in the analysis, the moderator variable of subsequent filtering was significant, $Q = 6.79$, $df = 2$, $p = .03$, suggesting that subsequent filtering contributed to the between-study heterogeneity found during the meta-analysis. The other moderators were non-significant, therefore suggesting they did not contribute to the heterogeneity (see *Table 4*). Additionally, a forest plot depicting the results for each moderator examined in the subgroup analysis can be seen in *Appendix B*. However, when the Wells et al. (1993) outlier was removed the subgroup analysis found that three moderator variables were significant, and each could explain the between-study heterogeneity found in the meta-analysis conducted for correct suspect identifications for target present lineups. The moderator variable of MTD type was found to be significant, $Q = 7.29$, $df = 1$, $p = .01$, subsequent filtering was also significant, $Q = 7.48$, $df = 2$, $p = .02$, as was the variable of participant type, $Q = 7.29$, $df = 1$, $p = .01$. The remaining two moderators were non-significant and are reported in *Table 5* below. All findings of the subgroup analyses can also be seen in the forest plot found in *Appendix B*.

Table 4

Test of Between-Subgroup Homogeneity for Correct IDs in Target Present Lineups

Moderators Examined	Findings
Pre-2012 vs post-2012	$Q = 0.01$, $df = 1$, $p = 0.93$
MTD type (Hybrid MTD vs pure MTD)	$Q = 0.82$, $df = 1$, $p = 0.36$
*Subsequent Filtering (Unclear, yes, no)	$Q = 6.79$, $df = 2$, $p = 0.03$
Participant Type (Community volunteer vs university student)	$Q = 0.82$, $df = 1$, $p = 0.36$
Study Type (Online vs in-person)	$Q = 0.01$, $df = 1$, $p = 0.93$

Note: Wells (1993) outlier included in this table

¹ While there was no evidence of heterogeneity for the other decision outcomes, moderator analyses were conducted across all outcomes for completeness. The results of these analyses are reported in *Appendices C-J*.

Table 5

Test of Between-Subgroup Homogeneity for Correct IDs in Target Present Lineups

Moderators Examined	Findings
Pre-2012 vs post-2012	$Q = 0.49, df = 1, p = 0.49$
*MTD type (Hybrid MTD vs pure MTD)	$Q = 7.29, df = 1, p = 0.01$
*Subsequent Filtering (Unclear, yes, no)	$Q = 7.48, df = 2, p = 0.02$
*Participant Type (Community volunteer vs university student)	$Q = 7.29, df = 1, p = 0.01$
Study Type (Online vs in-person)	$Q = 0.49, df = 1, p = 0.49$

Note: Wells (1993) outlier not included in this table

The subgroup analysis for incorrect filler identifications in target present lineups found that when the Wells et al. (1993) outlier was included in the analysis, the only moderator variable that was significant was subsequent filtering, $Q = 8.38, df = 2, p = .02$, suggesting that this variable may have contributed to the heterogeneity. The remaining moderator variables, which were found to be non-significant, are reported in *Table 6* below. Additionally, a forest plot depicting the results for each moderator examined in the subgroup analysis can be seen in *Appendix C*. When the Wells et al. (1993) outlier was removed, the subgroup analysis found that three moderator variables were significant and could explain the between-study heterogeneity found for incorrect filler identifications for target present lineups. The moderator variable MTD type was found to be significant, $Q = 8.12, df = 1, p < .01$, subsequent filtering was also significant, $Q = 8.88, df = 2, p = .01$, as was the variable of participant type, $Q = 8.12, df = 1, p < .01$. The remaining moderators were nonsignificant and are reported in *Table 7* below. All findings of the subgroup analyses can also be seen in the forest plot found in *Appendix C*.

Table 6

Test of Between-Subgroup Homogeneity for Filler IDs in Target Present Lineups

Moderators Examined	Findings
Pre-2012 vs post-2012	$Q = 0.02, df = 1, p = 0.88$
MTD type (Hybrid MTD vs pure MTD)	$Q = 2.54, df = 1, p = 0.11$
*Subsequent Filtering (Unclear, yes, no)	$Q = 8.38, df = 2, p = 0.02$
Participant Type (Community volunteer vs university student)	$Q = 2.54, df = 1, p = 0.11$
Study Type (Online vs in-person)	$Q = 0.02, df = 1, p = 0.88$

Note: Wells (1993) outlier included in this table

Table 7

Test of Between-Subgroup Homogeneity for Filler IDs in Target Present Lineups

Moderators Examined	Findings
Pre-2012 vs post-2012	$Q = 0.17, df = 1, p = 0.68$
*MTD type (Hybrid MTD vs pure MTD)	$Q = 8.12, df = 1, p < .001$
*Subsequent Filtering (Unclear, yes, no)	$Q = 8.88, df = 2, p = 0.01$
*Participant Type (Community volunteer vs university student)	$Q = 8.12, df = 1, p < .001$
Study Type (Online vs in-person)	$Q = 0.17, df = 1, p = 0.68$

Note: Wells (1993) outlier not included in this table

Choosing

Choosing refers to the rate at which lineup members were selected from the lineup.

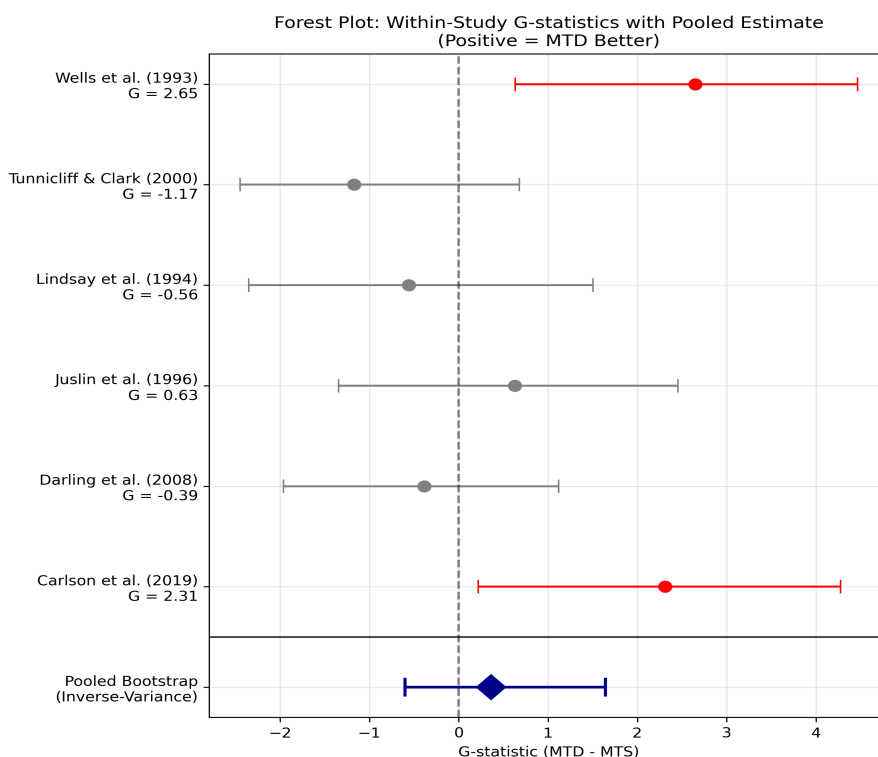
Collapsed across target presence, the analysis indicated the odds of choosing were significantly higher for MTD lineups compared to MTS lineups, $\ln OR = 0.42 [0.25, 0.58], p < .001$. In other words, eyewitnesses were approximately 1.52 [1.29, 1.80] times more likely to make an identification from a lineup created using the MTD method than they were from a lineup created using the MTS method. This suggests that lineups created using the MTD method increase choosing behaviour from eyewitnesses relative to MTS lineups.

Discriminability (d')

The results of the discriminability analyses showed the overall weighted effect size did not differ significantly from zero. The weighted meta-analytical mean suggests there is no significant difference in discriminability between the two lineup construction methods, $G = 0.36$ $[-0.60, 1.64]$, $p > .05$. As seen in *Figure 8*, the weighted mean is positive but has confidence intervals that cross zero. This finding highlights the conflicting information regarding which lineup construction method is the superior. Though two of six studies found the MTD method significantly increased discriminability compared to the MTS method, the remaining studies found no significant difference between the two methods (*Figure 8*). As so few studies met the criteria for the meta-analysis, more research is needed to investigate how lineup construction method impacts discriminability.

Figure 8

Forest Plot Depicting G-Statistics and Weighted Mean



While the overall weighted effect size found no significant differences in discriminability between the two lineup construction methods, it is also worth looking deeper at the individual study level data. This data suggests the MTD method may produce lineups which facilitate a witness' ability to distinguish between guilty and innocent suspects. As seen in *Table 8*, three studies show larger d' scores for MTD lineups compared to MTS lineups, while the other three studies show the opposite. This finding highlights the conflicted nature of the literature as to which filler selection method leads to better lineup outcomes.

However, with that said, two of the six studies in this analysis show the MTD method led to significantly better discriminability compared to the MTS method as indicated by a G-statistic greater than zero. Both Carlson et al. (2019), $G = 2.31 [0.22, 4.27]$, $p < .05$, and Wells et al. (1993), $G = 2.65 [0.63, 4.47]$, $p < .05$, showed lineups created using the MTD method significantly improved a witness' ability to distinguish between guilty and innocent suspects compared to lineups created using the MTS method. This finding suggests the MTD method of lineup construction may improve discriminability which is important in reducing eyewitness misidentifications. With that said, the four remaining studies found no significant difference in discriminability between the two lineup construction methods. It is also worth noting that no studies showed a significant improvement in discriminability by lineups created using the MTS method. Thus, further research is needed to determine which lineup construction method leads to better discriminability and how lineup construction method may indirectly impact lineup outcomes.

Table 8*Discriminability Statistics for MTS and MTD Lineups*

Study ID	MTS (d')	MTD (d')	G- statistic	95%CI Lower	95%CI Upper	Significant at .05
Carlson et al. (2019)	0.671	0.99	2.31	0.217	4.274	True
Darling et al. (2008)	1.899	1.614	-0.386	-1.961	1.119	False
Juslin et al. (1996)	1.161	1.37	0.628	-1.347	2.453	False
Lindsay et al. (1994)	1.782	1.492	-0.561	-2.353	1.502	False
Tunnicliff & Clark (2000)	1.941	1.229	-1.171	-2.448	0.679	False
Wells et al. (1993)	0.388	1.61	2.646	0.633	4.467	True

Discussion

The purpose of the meta-analysis was to answer the following question: does the MTD method improve lineup outcomes compared to the MTS method of lineup construction? The results of the meta-analysis suggest there may be benefits to using the MTD method of lineup construction when the target is present. For instance, the results show witnesses have higher odds of making a correct identification from lineups constructed using the MTD method compared to lineups created using the MTS method. However, when the target is absent the findings appear to be mixed. The results of the meta-analysis appear to suggest that when the target is absent from the lineup, witnesses are more likely to make an innocent suspect identification from MTD lineups than from MTS lineups. Though it should be noted that this finding is being driven by a single, heavily weighted study. On the other hand, the analyses for filler identifications and correct rejections found no significant differences between the two lineup construction methods. These findings suggest there may be a benefit to using the MTD method when the guilty perpetrator is present in the lineup, though it is clear additional research is needed to investigate how lineup construction method impacts witness identification decisions when the lineup contains an innocent suspect.

Choosing and discriminability analyses offered important insight into how lineup construction method impacts choosing behaviour and witness discriminability. The results show choosing, or witnesses identifying a member of the lineup, is higher in MTD lineups than MTS lineups. This suggests that lineups created using the MTD method increase choosing behaviour from eyewitnesses relative to lineups created using the MTS method. Additionally, discriminability analyses at the individual study level showed the MTD method of lineup construction may improve discriminability. This finding offers support for the speculation by Luus and Wells (1991) that the MTD method allows for more variation among lineup members, or propitious heterogeneity, which in turn leads to increased discriminability (Wells et al., 1993; Wixted & Mickes, 2014). However, at the group comparison level, the weighted meta-analytic mean suggested no significant differences in discriminability between the MTD and MTS methods of lineup construction. These contradictory findings may be due to differences in the ways in which the MTD method of lineup construction has been implemented in research.

As previously discussed in the systematic review, a lack of transparency in reporting made it difficult to determine exactly how the MTD method was operationalized in these studies. Therefore, it is possible MTD lineup constructors may not have been blinded to the suspect's appearance which may have influenced their filler selections. As research has shown how important blinding is in the administration of suspect lineups to prevent administrator biases from influencing witness identification decisions, it stands to reason that lineup constructors using the MTD method should also be blinded to ensure their filler selections remain unbiased. Thus, since not all MTD lineup constructors included in this meta-analysis were blinded to the suspect's appearance, it could have resulted in MTD lineup constructors selecting fillers who resemble the suspect more than they would if only relying on the description. This increase in

suspect-filler similarity could mimic that seen in MTS lineups, which would in turn lead to decreases in discriminability for MTD lineups (see DFD theory, Wixted & Mickes, 2014).

This meta-analysis demonstrates that there are very few studies which have directly compared the MTD method of selecting fillers to that of the MTS method. As such, it highlights the importance of conducting further research on this area to determine which lineup construction method yields better quality lineups and improves lineup outcomes. This research could have important implications for law enforcement. In the UK, there are currently no set guidelines for constructing lineups, with the only advice being that fillers should resemble the suspect based on physical characteristics such as race, gender, approximate age, build, and hair colour to prevent the suspect from standing out in the lineup (Police and Criminal Evidence Act 1984, Code D, 2023). Additionally, as eyewitness misidentifications continue to play a role in nearly a quarter of all wrongful convictions, it is imperative law enforcement are using lineup procedures with the best possible outcomes.

Limitations

While the findings of this meta-analysis offer some support for the use of the MTD method of lineup construction, it does have its limitations. One important limitation is that only seven studies met the inclusion criteria for meta-analysis which may have resulted in low statistical power, especially for the moderator analyses. In addition to including a small number of studies, the majority of those included were before the replication crisis in 2012. This means five of the seven studies included in the meta-analysis were underpowered and did not contribute much to the weighted means. As such, the remaining two studies in the analysis contained large sample sizes, leading to them being weighted more heavily than the others and potentially skewing the results of the meta-analysis. Therefore, the findings of this analysis should be taken

with caution. With that said, this meta-analysis highlights the issue that there has been little research directly comparing the two methods of lineup construction.

Future Research

The results of this meta-analysis show that the research on how lineup construction method impacts identification decisions is conflicted. It also revealed there are gaps in the literature, such as how blinding of MTD lineup constructors impacts the effectiveness of the method. Therefore, this systematic review and meta-analysis informed the study design for my doctoral research. In this research, we aim to further investigate which lineup construction method, the MTS or MTD method, leads to better lineup outcomes. We also aim to investigate whether blinding MTD lineup constructors improves the effectiveness of the method. Lineup constructors will be provided with either an image of the suspect (MTS method), a perpetrator description (blinded MTD method), or both an image of the suspect and the matching perpetrator description (unblinded MTD method) and instructed to select filler who best match the information provided. Across the subsequent chapters of this thesis, those lineups will be evaluated for lineup fairness, suspect-filler similarity, goodness of fit to the description, and identification outcomes.

Chapter 2

Eliciting person descriptions and constructing suspect lineups

While much research has been conducted on the creation and administration of suspect lineups, the research is often conflicting as to which method of filler selection leads to better lineup outcomes, such as witness accuracy, confidence, and lineup fairness (see Chapter 1). This chapter describes two studies that were designed to create lineups using different filler selection methods, which could then serve as stimuli in the subsequent chapters of this thesis. This study directly compares the commonly used MTS method of filler selection to two variations of the MTD method to investigate how filler selection methods influence the quality of fillers selected. In one variation of the MTD method, lineup constructors will rely solely on the perpetrator's description (blinded MTD) when selecting fillers. In the other variation, lineup constructors will have access to an image of the suspect alongside the description when selecting fillers, though they will be instructed to base their selections on the description (unblinded MTD). In order to do this, we conducted two studies. The aim of the first study (1a) is to elicit perpetrator descriptions of a sufficient quality that they can be used to create the MTD lineups in the second study (1b). Once the lineups are created, they will be evaluated in a series of experiments to determine which filler selection leads to fairer lineups, as well as improved witness accuracy and confidence.

Constructing Suspect Lineups

Suspect lineups continue to play a critical role in eyewitness identifications, though research has shown that lineups can negatively impact eyewitness identification accuracy if not created or administered properly. As a result, much research has been conducted in efforts to inform law enforcement officials of the best ways to create and administer “fair” lineups. Wells

and colleagues (2020) laid out a series of recommendations for ensuring lineups are fair, including: providing witnesses with unbiased pre-lineup instructions, administering a double-blind lineup (i.e. both lineup administrators and witnesses do not know who the police suspect is), and selecting adequate fillers, or known-innocent individuals, to prevent the suspect from standing out. These recommendations also state that lineups should consist of only *one* suspected individual and at least *five* fillers. While the five filler minimum recommendation is somewhat arbitrary, the use of known-innocent fillers allows for unreliable witnesses to mistakenly choose a filler rather than an innocent suspect; therefore, protecting an innocent suspect from mistaken identification (Wells et al., 2020). These lineups should then be presented to the eyewitness in an acceptable format such as live, in-person, photographic, or video.

While six-person lineups (i.e. one suspect along with five fillers) are the recommended minimum for lineup constructors, there is considerable heterogeneity in lineup sizes used in research and in practice by law enforcement agencies. Research conducted by Fitzgerald, Rubínová, and Juncu (2021) found lineups can range in size from 3- to 10-person lineups, with several jurisdictions and countries around the world utilising 8-10 person lineups. However, the number of fillers present, or nominal size of the lineup, does not matter as much as ensuring that all of the fillers included in the lineup are ‘quality’ fillers. Fillers are considered to be quality fillers if they prevent the suspect from standing out among them. It is therefore recommended the fillers closely match the culprit description provided by the eyewitness, and resemble the suspect, to prevent the suspect from standing out (Wells et al., 2020). Luus and Wells (1991) argued quality fillers play an important role in reducing eyewitness misidentifications.

The use of quality fillers helps provide protection for innocent suspects in several ways. One way in which fillers help reduce misidentifications is by allowing for known errors (Luus &

Wells, 1991). For instance, if police administer a lineup containing only potential suspects, then any identification would be considered incriminating evidence against the person identified. However, if police provide a lineup in line with recommendations (i.e. containing one suspect and several fillers), then a witness identification of a filler would be considered an error, rather than incriminating evidence. Allowing for known errors not only protects innocent suspects but can also demonstrate to police that the witness's recognition memory of the perpetrator is fallible (Luus & Wells, 1991). The presence of quality fillers also helps control for guessing from witnesses who are uncertain. Even when eyewitnesses are provided with pre-lineup instructions stating the suspect may not be present in the lineup, witnesses may still feel pressured to make an identification. In instances where the witness is simply guessing, the chance of selecting the suspect is reduced to $1/N$ (N being the nominal size of the lineup) (Luus & Wells, 1991; Colloff et al., 2021). Another way in which the presence of quality fillers helps reduce misidentifications is by ensuring the lineups are testing the witness' recognition memory for the perpetrator rather than recall (Luus & Wells, 1991). Decades of research on cognition and memory have argued that recall and recognition involve different cognitive processes (Robinson & Johnson, 1996). Recall is often characterised by the intentional and effortful retrieval of information and is tested when a witness is providing a description of the perpetrator to police (Robinson & Johnson, 1996; Luus & Wells, 1991). Recognition, on the other hand, tends to be a less effortful process as it is based on familiarity heuristics and can be tested when a witness is provided a suspect lineup to decide whether the lineup member are familiar or unfamiliar (Robinson & Johnson, 1996; Rotello, 2000; Luus & Wells, 1991). While there is agreement that quality fillers are needed for a lineup to be considered fair, there is conflicting information on how those fillers should be selected.

There are two commonly used methods of filler selection used to create lineups, the previously mentioned MTS method and MTD method. While the MTS method is more widely used and selects fillers based on their physical similarity to the suspect, the MTD method uses perpetrator descriptions elicited by law enforcement to base filler selections on (Lindsay et al., 1994; Malpass, Tredoux, & McQuiston-Surrett, 2007; Luus & Wells, 1991; Wells et al., 1993). The MTD method was originally proposed in response to a problem with too-high lineup member similarity, sometimes seen in MTS lineups, as it has a natural stopping point in which characteristics fillers should match the suspect on. By selecting fillers who match the description, it allows the fillers to be similar enough to prevent the suspect from standing out yet still provide variability which allows the witness to distinguish between the individuals in the lineup (Wells et al., 1993). As discussed in Chapter 1, Carlson and colleagues (2019) found that the MTD method of filler selection yields lineups with higher amounts of “propitious heterogeneity” compared to the MTS method, meaning the MTD method was more likely to lead to heterogeneous lineups which allows for the fillers and suspects to look similar to one another but still also allows for discriminability among lineup members (Carlson et al., 2019; Wixted & Mickes, 2014). This helps the witness distinguish between the innocent and guilty, as well as be more accurate in their decisions.

This finding was further supported by the results of the meta-analysis conducted in Chapter 1, which found the odds of a correct identification is higher in lineups created using the MTD method compared to lineups created using the MTS method. However, the meta-analysis also showed that the odds of an innocent suspect identification were also higher for MTD lineups than for MTS lineups. Though it is worth noting this finding was driven by a single, heavily weighted study which may have skewed the results of the impact of filler selection method on

innocent suspect identifications. Additionally, the meta-analysis revealed filler identifications were more likely to occur in MTS lineups than MTD lineups when the target was present. There were no differences between the two methods regarding non-identifications for target present or target absent lineups. Thus, the findings of the meta-analysis provided support for the MTD method when the guilty perpetrator was present in the lineup but suggests the research remains conflicted as to which method leads to better outcomes when the suspect is innocent. While research has shown that the MTD method can yield quality lineups, it has been implemented in various ways, which may impact the method's effectiveness.

The implementation of the MTD method has varied significantly since it was originally proposed by Wells et al. (1993), where the lineup constructor had no knowledge of the suspect, resulting in fillers who were selected solely based on the culprit description. As discussed in Chapter 1, a review of the existing research shows that the most common variation in how this method is operationalised is that the lineup constructor has existing knowledge of the suspect's appearance in some capacity. For example, fillers are often initially selected from a large pool of options using the culprit description. Then a second round of filtering takes place in which those who best resemble the suspect are selected for the final lineup. One example of this includes evaluating each of the fillers from the description-narrowed pool for physical similarity to the suspect and selecting those with higher similarity ratings for the final lineup (Carlson, Gronlund, & Clark, 2008; Lucas, Brewer, & Palmer, 2021). Another variation uses similarity ratings in a two-part procedure to, first, compare the similarity of the narrowed pool of fillers to the culprit description and, secondly, compare the lineup members to one another (Koehnken, Malpass, & Wogalter, 1996; Haw & Fisher, 2004). Once ratings for similarity to the description and similarity to one another have been obtained, those fillers with the highest ratings for both ratings

make up the final lineup (Haw & Fisher, 2004).

However, these variations in how the MTD method is implemented may lead to lineup constructors who are biased in their filler selections, which may undermine the effectiveness of the MTD method. If a lineup constructor has existing knowledge of the suspect's appearance, they may be biased in their filler selections, as top-down information processing dictates that prior knowledge and experience influences perception and decision making (Firestone & Scholl, 2014; Kinchla & Wolfe, 1979). Given that experience, prior knowledge, and/or perspective can impact the views and decisions of an individual, it seems inevitable that they will also influence the construction and administration of lineups. For example, imagine you are a lineup constructor using the MTD method and the description includes the perpetrator had blonde hair. Blonde hair covers a broad range of shades. Now imagine during the filler selection process, you saw an image of the suspect and noticed that their blonde hair is a very light shade. Having this knowledge may bias you towards selecting fillers with lighter blonde hair, despite this not being stipulated in the description. Thus, reducing the potential variability among the fillers selected compared to those that may have otherwise been selected if you had not seen an image of the suspect. Therefore, while much research has been conducted on lineup administration and blinding of the suspect's identity, little is known about how prior knowledge of a suspect's appearance may bias the constructor's selection of fillers and undermine the effectiveness of the MTD method.

Importance of Description Quality on the Match-to-Description Method

While the MTD method of filler selection has the potential to yield better quality lineups compared to MTS lineups, the success of the method relies heavily on law enforcement eliciting a detailed description of the perpetrator. Research has shown that vague descriptions can impact

the success of the MTD method by resulting in biased lineups, which can be especially problematic when the suspect is innocent (Lindsay et al., 1994). Lindsay and colleagues (1994) suggest that the less detail included in the perpetrator description, the more likely the lineup created would be biased towards the suspect. This is due to the suspected individual having higher chances of matching the perpetrator on details that may have been omitted from the description compared to the fillers selected. Take the following witness description for example, “young, White male, brown hair, average build”. This description lacks information regarding height, hair length and style, facial features, as well as any potentially unique features (i.e. moles, scars, tattoos). As a result, lineup constructors could select fillers who match the features mentioned in the description but vary drastically from one another regarding height, hair length, and facial hair. If the suspected individual is the only lineup member with short, cropped hair, they may be more likely to match the memory the witness has of the perpetrator than the other lineup members who have slightly longer brown hair. As such, it is imperative that law enforcement officers elicit quality, detailed descriptions from witnesses.

Description Elicitation

Eliciting accurate, detailed person descriptions is critical for police investigations. It is especially important for narrowing down a suspect pool and for constructing police lineups for eyewitness identification, specifically when using the MTD method of lineup construction (Milne & Bull, 2006; Dalton, 2021). As the MTD method of lineup construction depends on an accurate and detailed witness description of the culprit to base filler selection on, it is necessary to find efficient and cost-effective ways for police to elicit these descriptions (Luus & Wells 1991; Wells et al., 1993). With that said, other than prompting witnesses to provide information regarding a perpetrator’s race, age, gender, and build, there are currently no set procedures in

place for police to obtain complete and detailed person descriptions (Dalton, 2021). This lack of procedural guidance often leads to police obtaining vague person descriptions despite the existing research on how to elicit reliable information from witnesses.

Decades of research has shown that regardless of improvements in interviewing techniques, witnesses often provide vague person descriptions of perpetrators. Lindsay and colleagues (1994) found that person descriptions tend to be vague and lacking sufficient detail, with witnesses reporting on average only 7.35 physical, whole person characteristics or details. This finding by Lindsay et al. (1994) is supported by archival studies which have shown witnesses average anywhere from 3.9 to 9.7 physical descriptors when describing a perpetrator (Dalton, 2021). Additional research has shown the information most frequently reported by witnesses includes descriptions of clothing, hair colour, and height; while specific facial features of the perpetrator are rarely mentioned (Lindsay et al., 1994; Demarchi & Py, 2009). Lindsay et al. (1994) also found even when details were reported they still tended to be vague generalizations or ranges, such as they were “tall”, or they were in their late 20s to mid-30s. Additionally, they found certain characteristics of a perpetrator may be omitted from a description when that characteristic of the culprit, such as race or gender, falls into what would be considered a community majority or the crime committed is generally committed by one gender, and therefore seems “obvious” to the witness and even law enforcement. These “obvious” characteristics are known as *default values* and are often accidentally omitted by witnesses but are important pieces of information that must be gathered to have a complete description of the perpetrator.

Default values and vague descriptions could be problematic as these descriptions are then used as the inclusion criteria for fillers included in a police lineup. If the description is too vague,

this could result in a description-matched lineup that is biased towards the suspect, as there may potentially be too much variability between the suspect and the fillers. For example, imagine the following hypothetical witness description: “male, approximately mid-20s, having short dark hair, dark eyes, and of average build”. This description does not include race, because the witness assumed that race was obvious (i.e., they viewed ‘White’ as the default). A lineup constructor could, therefore, include fillers from different racial backgrounds, thereby reducing the effective size of the lineup. While this is an extreme example, it demonstrates how important it is for police officers to elicit complete and detailed person descriptions from witnesses. Therefore, investigating ways in which to elicit detailed information from witnesses is needed, as is testing these descriptions by using them in the creation of suspect lineups.

Research has shown that there are several variables which could influence the accuracy of a person description (Brown, LLOYD-Jones, & Robinson 2008). These influencing variables can be separated into estimator and system variables, estimator variables being those that are not within the control of the criminal justice system and include witness characteristics, characteristics of the crime or event which could affect memory encoding, as well as characteristics of the testimony (Wells & Olson, 2003). For example, witness characteristics, such as race, have been shown to influence facial memory and person descriptions from witnesses when the perpetrator is of a different race from them (Meissner & Brigham, 2001; Brigham et al., 2007). While race can play an influential role in witnesses reporting detailed person descriptions, encoding variables such as witness intoxication, lighting, and distance from the perpetrator can also be important variables which impact witness accuracy and completeness in reporting perpetrator descriptions (Wells & Olson, 2003; Meissner, Sporer, & Schooler, 2014).

Additionally, witnesses may provide poor descriptions, at least in part, due to a discrepancy between the process of encoding and recognizing a face and the process of recalling and verbally describing the features of the face (Breznitz, 2002; Dalton, 2021). This discrepancy arises as the processes of encoding visual information and recognizing faces use a holistic approach, while the process of recalling and describing faces tends to be featural, or sequential, and asks witnesses to describe faces as parts instead of as the whole in which they were initially encoded (Tanaka & Farah, 1993; Gabbert & Brown, 2015; Kask & Bull, 2009). Therefore, this difference in processing information makes it difficult for witnesses to properly verbalise the physical characteristics of perpetrators (Breznitz, 2002; Kask & Bull, 2009). However, it is important to note that this not only affects how witnesses describe features of the perpetrator, but also how the police officers then interpret and visualize the verbal descriptions provided. For example, if the witness describes the perpetrator as having a large nose or as having “droopy eyes” there are any number of ways for the officer to interpret that information coloured by their personal experiences (Kask & Bull, 2009).

Interviewing Techniques to Elicit Person Descriptions

Research on interviewing techniques has shown that solely using closed-ended questions (i.e. questions resulting in “yes” or “no” answers) is less effective at eliciting detailed information from witnesses compared to using a combination of open-ended and closed-ended questions (Meissner, Sporer, & Schooler, 2014). Methods such as free recall, the Cognitive Interview (CI), the Person Description Interview (PDI), and more recently the Self-Administered Interview (SAI) have been researched and developed to gather details from witnesses and improve recall (Geiselman et al., 1984; Demarchi & Py, 2009; Gabbert, Hope, & Fisher, 2009). Free recall often uses a series of open-ended questions, such as “what can you remember about

the person or event?”, to elicit details of an event and those involved, followed up by closed questions to gather any additional information missing from the initial free recall response. While more recent research has shown that free recall responses can be lacking in sufficient detail to properly narrow a suspect pool, this method is still used by law enforcement officers (Lindsay et al., 1994; Brown et al., 2008). Brown and colleagues (2008) found that when law enforcement officers were asked which interviewing techniques they used to elicit person descriptions in the field, the vast majority answered that they tend to use an open-ended question to encourage free recall in the witness. Additionally, less than half of those law enforcement officers state that they then used a series of specific probes to gather additional information on general physical characteristics such as race, gender, build, or haircut and colour, with even fewer law enforcement officers asking for more detailed information regarding specific features like facial features, facial hair, piercings, tattoos, or other distinguishing features.

This lack of detail which free recall methods can yield led to the development and proposal of the Cognitive Interview (CI). The CI is a well-established interviewing procedure commonly used by police during an investigation to gather reliable information from witnesses. However, this technique tends to be very time extensive for law enforcement and focuses more on the event, or the “what”, “when”, and “where” questions of a criminal investigation (Geiselman et al., 1984; Brown et al., 2008). With that said, certain principles of this interviewing procedure such as the change of order and reporting of all information have been utilized in the Person Description Interview (PDI) to elicit person descriptions (Geiselman et al., 1984; Demarchi & Py, 2009).

The Person Description Interview (PDI) was developed to elicit detailed person descriptions from witnesses in a two-step interview (Demarchi & Py, 2009). The first step

involves a “general-to-specific instruction” which was designed to enhance the descriptive strategy naturally used by people to describe others. In this step, witnesses are instructed to describe the individual as completely as possible through the use of both general details (e.g. height, build, ethnicity) and featural details (e.g. facial features, identifying tattoos or scars). The second step involves a “down-to-top instruction” which utilizes the change of order principle from the CI (Geiselman et al., 1984; Demarchi & Py, 2009). The use of the change of order principle goes against the natural “top-to-bottom” description strategy often used by people and, therefore, enhances the recall of lower facial features. In this step, witnesses are instructed to describe the face of the individual by starting at the bottom of the face (e.g. chin and mouth) and moving upwards. A field study tested the PDI against the *standard French police interview for person description* (SFPIPD), an interviewing procedure that uses a free recall prompt, followed by questions about any physical traits not mentioned by witnesses. This study found the PDI outperformed the SFPIPD in terms of completeness and accuracy, suggesting the PDI elicits more accurate and detailed person descriptions than free recall techniques used by police.

The Self-Administered Interview (SAI) has also successfully integrated and utilized principles of the CI, such as context reinforcement and reporting of all information, to elicit detailed person descriptions (Geiselman et al., 1984; Brown et al., 2008; Gabbert et al., 2009). The SAI was developed to elicit thorough and detailed statements from witnesses by having them record their memories of the event on their own, while still following a set protocol including instructions and questions (Gabbert et al., 2009). This investigative tool is divided into five sections, with one of the sections specifically focusing on eliciting person descriptions. Instructions, based on the key components of the CI which are mental reinstatement of context and reporting every detail, are provided to aid witnesses in memory retrieval. These instructions

tell witnesses to picture the individual in their mind, as well as provide them with a list of memory cues to help witnesses visualize the individual in the most detail possible. Witnesses are then asked to provide a complete and accurate description of the person, while also telling them not to guess about features or characteristics they are not sure of. In combination with the other sections, the SAI aids eyewitnesses in providing detailed accounts of the people and events witnessed. As such, several studies have shown the SAI yields accurate information at similar rates to an interviewer-administered CI, all the while reducing the strain on police resources (Gabbert et al., 2009; Hope et al., 2011; Gabbert et al., 2012).

Since the proposal and implementation of the SAI, several versions have been created to address specific incidents such as missing persons (SAI-MISSING), workplace incidents (SAW-IT Workplace Incidents), and road traffic incidents (SAI-RTC) in efforts to aid law enforcement while gathering the necessary information each type of incident requires (Gabbert et al., 2020; MacLean, 2019; Horry et al., 2022; Self-Administered Interview, 2021). Like with the original SAI, the SAI-MISSING was created to elicit detailed person descriptions to aid in the search for missing persons (Gabbert et al., 2020). This tool is broken into three sections, the first focuses on eliciting a detailed and accurate physical description of the missing person and uses mental reinstatement of context instructions to help individuals provide as much information about the missing person as possible, the second section focusing on the circumstances surrounding their disappearance, and the third section is used to gather information about the missing person's normal routines and behaviours and any potential changes to these characteristics. Gabbert and colleagues (2020) found that techniques used in the SAI-MISSING helped individuals create their own memory cues, which resulted in more detailed descriptions of the missing person being provided compared to a self-administered control form. While these new techniques are helpful

tools to elicit high-quality person descriptions, more research is needed to determine if these methods consistently yield better person descriptions than other elicitation methods. This research is important as the successful construction of fair suspect lineups may depend on these descriptions elicited by law enforcement.

Study Aims

This study is split into two parts, with separate aims. Study 1a aimed to elicit high-quality and low-quality person descriptions that are suitable for lineup construction. To this end, participants watched either high- or low-quality mock crime videos and then described the perpetrator using either a modified version of the *Missing Person Self-Administered Interview* (SAI-MISSING) or a free recall form. To solve the potential issue of participants omitting default values accidentally, both description elicitation methods were supplemented with a default values checklist.

Study 1b aimed to investigate how blinding the lineup constructor to the suspect's appearance affects filler selections and filler selection strategies. Lineup constructors were either provided only with the descriptions elicited in Study 1a (blinded MTD condition), with only the suspect's image (MTS condition), or with both the description of the perpetrator and the suspect's image (unblinded MTD condition).

While the primary aim for the first part of this study was to elicit descriptions varying in quality to be used in the creation of suspect lineups, we hypothesised that descriptions would include more unique descriptors in the high-quality, SAI, condition than in the low-quality, free-recall, condition as encoding (or video) quality was purposefully confounded with description elicitation method. Additionally, while the focus of the second part of the study was to create the lineups that will be used in subsequent studies, we predicted there would be little overlap in

fillers selected for lineups using the blinded version of the MTD method with those selected using the MTS method. We also predicted that the unblinded version of the MTD method would have more fillers in common with MTS lineups than with the blinded MTD lineups. We also predicted that lineup constructors in the unblinded MTD condition would report basing their filler selections, at least in part, on the suspect's image.

Method 1a

Participants. A sample of 15 Swansea University students were recruited to participate in this study. These individuals were recruited through the online university participant pool and given course credit in exchange for participation. Due to an oversight, participant demographics were not collected for this study. Sample size was determined by looking at past lineup research which created a modal description to use in filler selection and mock witness tests. Based on that research, it was determined that a minimum of three witness descriptions were needed per stimulus to provide sufficient detail for the modal description, which necessitated 15 participants.

Design. The experiment followed a between-subjects design. Participants were randomly assigned to a description quality condition, high-quality ($n=7$) or low-quality ($n=8$), and were asked to provide descriptions for three perpetrators, which were randomly selected from a pool of six. Those in the high-quality description condition were provided with higher quality mock crime videos, which should allow for better encoding of the perpetrator's appearance and then prompted with a modified version of the SAI to elicit their description of the perpetrator. Those in the low-quality condition, however, were provided a lower quality mock crime video, and then given a free recall prompt to elicit their descriptions of the perpetrator. Within each condition, each stimulus was described by 3-4 participants. Note that elicitation method and encoding quality were purposely confounded to elicit richer and sparser descriptions to test in future

studies.

Materials

Mock Crime Videos. The staged mock crime videos used in this study came from research conducted by Molinaro et al (2021). These stimuli were used with permission and are openly available via a Creative Commons license (CC-BY 4.0), stimulus materials can be located through Open Science Framework

(https://osf.io/dz5pf/?view_only=2b41a6ce5c4d43e0abb08ea1f894d6b2). To gather various qualities of person descriptions to use in future research, we had both high-quality and low-quality staged, mock crime videos. These videos were the original materials from the Molinaro et al (2021) study and were not edited beyond what was done for that study. The high-quality videos were just over 1-minute in length and were filmed using a tight, or close-up, view of the crime and the perpetrator. These videos showed the perpetrator's face for a full 60 seconds. The low-quality videos were approximately 15 seconds in length and were filmed from further away, using a wide view of the perpetrator and the crime scene. These videos showed the perpetrator's face for only 10 seconds. Each mock crime video included a different perpetrator. These perpetrators were quite distinct and varied in age, gender, and overall physical appearance.

Description Elicitation. In the High-Quality condition, participants completed a modified version of the *Missing Person Self-Administered Interview*, (SAI-MISSING; Gabbert et al., 2020). For this study's purpose, we used only the third section of this tool, which focuses specifically on person descriptors (*Appendix L*). This section of the SAI has the individual envision the person they are wanting to provide a description of. Participants were given the following instructions, "Thank you for helping us with our efforts to identify the culprit caught on video. We would like to ask for some further information. The following form has been designed for you to complete

independently, so that you can take all the time you need to answer the questions. Any information you provide might prove incredibly useful in the search for the individual responsible. However, you are free to report as little or as much information as you wish. Additionally, do not worry if you do not have or know all of the information the form asks for.” They were then instructed to envision the perpetrator as follows, “Please envision the perpetrator seen in the video. Hold the image of this person in your mind, then read through the list of memory cues below to see if they help you remember information that we will find useful.” Once they had envisioned this person, they were instructed to read through a list of physical characteristics which acted as memory cues. The memory cues provided included characteristics like body build, hair colour, hair length or style, and facial appearance and complexion. After they had done this, they were instructed to provide a detailed physical description of the perpetrator, “Please now provide a detailed physical description of the perpetrator. Be as complete and accurate as possible. Please do not guess on any details which you are unsure.” They were then provided with an empty text box to type up their descriptions in.

For the low-quality person descriptions, participants were given a basic free-recall prompt. In this prompt participants were given the instructions as follows, “Please provide a detailed physical description of the perpetrator seen in the video. Be as complete and accurate as possible. Please do not guess on any details which you are unsure.” They were then provided with an empty text box to type up their descriptions in.

In both conditions, all participants were given a checklist to ensure no default values were omitted from the person description (*Appendix M*). Default values can include characteristics such as race/ethnicity, gender, age, and body type or build and as such, were included in the default values checklist.

Procedure. Participants first read a participant information sheet followed by an informed consent document. Once they had read both documents, they were asked if they agreed to participate. If they disagreed, they were sent to the end of survey page and allowed to exit out of the program. If they consented to participate, then they were then randomly assigned to either the high-quality description elicitation condition or the low-quality description elicitation condition.

Participants completed three trials – each corresponding to a different stimulus. In each trial, participants were instructed to watch a short, staged crime video before proceeding to the next screen. Once the video had ended, participants completed a 90-second distractor task (a word search puzzle). On the next screen, participants were asked to describe the perpetrator from the video that preceded the distractor task. The relevant prompts were displayed depending on condition (the modified SAI prompts or the free recall prompt) and participants typed their responses into a free text entry box displayed below the instructions. After providing their descriptions, participants were taken to the next screen which displayed the default values checklist. They were given a series of questions relating to race, gender, age, and body type and instructed to select the option that best fits the perpetrator seen in the video. Once they had completed the checklist, they were directed to the next staged mock crime video to complete another culprit description. After the participants submitted their final description, they were debriefed and thanked for their time. This study took approximately 15 minutes to complete.

Results 1a

This study utilized a modified version of the Self-Administered Interview (SAI), high-quality description elicitation method, and a basic free-recall questionnaire, low-quality description elicitation method, to deliberately confound questioning method with encoding quality and investigate how this impacted the amount of detail included in a perpetrator

description provided by a witness. We hypothesized the SAI would yield more descriptors being included in the perpetrator description as it utilizes cognitive interviewing strategies and memory clues to aid in witness memory.

For this study's purposes, a descriptor is any physical characteristic of the perpetrator mentioned by the witness (either in their written description or within the descriptor checklist) such as hair colour, length, physical build, face shape, or facial hair. Each physical characteristic was then tallied individually (e.g., hair colour and hair length are tallied as two separate characteristics). Take the following perpetrator description, for example: “white female, average height (approximately 5'7" to 5'8"), average build, short, brown bob-styled hair, approximately mid 50s to early 60s.” This description contains *eight* physical descriptors, the first descriptor being the race of the perpetrator, the second descriptor indicating their gender, then descriptions of approximate height and build were provided, followed by descriptors of hair length (short), hair colour (brown), and hair style (bob-styled), and lastly a description of approximate age was provided. The number of descriptors provided was then compared between those created through the SAI and from the free-recall questionnaire.

At first glance, looking at the descriptive statistics, it appears as though the high-quality description elicitation method yielded more descriptors of the suspect's physical appearance than the low-quality description elicitation method (*Table 9*). For five of the six different stimuli sets, witnesses provided more descriptors of the perpetrators in the high-quality description elicitation condition than in the low-quality description elicitation condition.

Table 9*Mean Number of Details in Person Descriptions by Stimulus and Quality Condition*

Stimulus	HQ Condition	LQ Condition
1	7.33 (2.08)	4.00 (1.41)
2	7.00 (1.73)	3.75 (2.22)
3	5.75 (2.63)	4.50 (0.58)
4	6.25 (0.50)	3.75 (1.71)
5	5.00 (1.73)	5.00 (0.82)
6	6.50 (3.11)	5.25 (1.26)

Note. Standard deviations for each mean are noted in parentheses.

A multi-level regression was conducted to investigate whether the knowledge of condition could predict whether an individual would provide more details in a culprit description. Conceptually, this analysis is similar to an independent t-test and can be interpreted in much the same way. This analysis was chosen as we have 45 observations that are not independent of each other as they share some important sources of variability. These shared sources include the observers and the stimuli, as the observations were nested within observers, who were likely to vary from each other in how much information they provide, and within stimuli, in which some stimuli likely elicited more descriptors than others. By conducting a multi-level regression over an independent t-test, we thereby reduce the likelihood of a Type 1 error. The analysis was conducted using the lme4 and lmerTest packages in R (Bates, Mächler, Bolker, & Walker, 2015; Kuznetsova, Brockhoff, & Christensen, 2017).

The outcome variable was the number of descriptors provided. The multi-level model included random intercepts for Participant and Stimulus; in other words, the baseline number of descriptions provided could vary between participants and between stimuli. The null model contained only these random intercepts. The null model was then compared to a model that included Condition as a fixed effect – with the High-Quality condition as the reference category.

Our hypothesis was that those in the high-quality description elicitation condition would provide more detailed culprit descriptions compared to those in the low-quality description elicitation condition. The model that included Condition was a significant improvement over the random effects model, $\chi^2(1, N = 45) = 6.25, p = .012$. The fixed effect of Description Quality, with High-Quality as the reference condition, was significant, $\beta = -1.91, SE = .74, t(12.23) = -2.59, p = .02$. The random intercept for Participant accounted for a substantial proportion of the total variance in number of descriptors provided, $\sigma^2 = 1.50, SD = 1.23, ICC = .48$, while Stimulus accounted for a smaller amount of the variance, $\sigma^2 = 0.09, SD = 0.30, ICC = .05$. Descriptions elicited in the Low-Quality condition included significantly fewer details ($M = 4.37, SE = 0.29$) than descriptions elicited in the High-Quality condition ($M = 6.29, SE = 0.44$), $t(43) = 3.71, p < .001$. This finding suggests we were successful in manipulating description quality.

The next step was to create modal descriptions for each stimulus. We applied a rule whereby a feature had to be included in at least two descriptions by separate witnesses to be incorporated into the modal description. While individuals in the high-quality description condition did provide more detailed descriptions compared to those in the low-quality condition, the overall modal descriptions for each culprit in both conditions vary only slightly (*Table 10*). This could be due to our modal description criteria, which was that for a detail to be included in the modal description, the detail had to be included by at least two participants. Thus, making it possible more descriptors would be counted across both conditions. These descriptions, as well as what was provided through the default value checklist, can be seen in the table below (*Table 10*). One can see how slight the differences are between the high- and low-quality descriptions for some of the stimuli. For example, for the first stimulus, the primary difference between the two descriptions is that the age was estimated in the high-quality description but not in the low-

quality description, as well as the mention that the individual has wrinkly skin which ties into the estimated age.

Table 10

Descriptions Elicited by Description Quality Condition

Stimulus	Default Value Checklist	LQ Condition	HQ Condition
1	Caucasian/White Male Approximately 55-64 Endomorph (heavy-set build)	White male, gray hair, stocky build, approximately in his 50s	Older white male, thinning, short gray hair, wrinkly skin, heavier set, late 50s to 60s
2	Caucasian/White Female Approximately 55-64 Ectomorph (lean and slender build) / Combination (average build)	White female, brown bob-styled hair, approximately in her 50s	White female, average height (approximately 5'7" to 5'8"), average build, short, brown bob-styled hair, approximately mid 50s to early 60s
3	Caucasian/White Male Approximately 35-44 Endomorph (heavy-set build) / Combination (average build)	White male, bald, stocky/bigger build, middle aged	White male, bald, has some facial hair/stubble, lean build, has tattoo on forearm, approximately mid 30s to early 40s
4	Caucasian/White Female Approximately 18-24 Endomorph (heavy-set build)	Young, white female, long brown hair, slightly overweight	Young, white or Latina female, taller than average height (approximately 5'10"), heavy set/larger build, mid to longer length dark hair, clear, blemish-free skin, early to mid 20s
5	Latino/Hispanic Male Approximately 45-54 Endomorph (heavy-set build) / Combination (average build)	Latino/ Hispanic male, short dark hair, tanned skin	Latino/Hispanic male, below average (male) height, black hair, clean shaven face, approximately mid 40s to early 50s
6	Caucasian/White Male Approximately 25-34 Endomorph (heavy-set build)	White male, approximately in his 30s, short brown hair, overweight/larger build	White male, shorter to average height, stocky/heavy build, short black hair, has short beard/stubble, rounded face, clear, blemish-free skin, mid 20s to early 30s

For this reason, we decided to proceed with the descriptions from the default value checklist (column 1) with the addition of the hair colour descriptor from the low-quality description (column 2) and the descriptions from the high-quality description condition (column 3) for the next study. This decision was made to ensure adequate differentiation between the two description conditions. As in real-world scenarios, police do not always elicit detailed, high-quality perpetrator descriptions from witnesses. In departments which utilise the MTD method to construct lineups, vague descriptions may still be used for filler selection. Therefore, to replicate plausible real-world scenarios and increase generalisability, descriptions varying in detail were provided to MTD lineup constructors for the filler selection process. By varying the quality of the descriptions used for lineup construction, it allows for variability within the MTD lineups that will be tested against several metrics.

Discussion 1a

We aimed to elicit high-quality and low-quality person descriptions that could be used to create MTD lineups in the next study. The results of this study show by purposely confounding the description elicitation method and the viewing condition (video quality), we were successful in manipulating the quality of perpetrator descriptions provided by eyewitnesses. However, these differences were quite minor and only differed by a couple of details, as even the descriptions provided in the low-quality conditions were surprisingly detailed. This finding could be due to several factors, such as that individual differences in detail may have been lost when combining each of the witness descriptions into one modal description for each perpetrator. Therefore, resulting in us needing to take a different approach for creating high-quality and low-quality descriptions suitable for inclusion in the next experiment.

Additionally, another factor which could have contributed to this small difference in

details reported could have been that the retention interval was short. Previous research on recognition and recall has shown that memory quality and accessibility decline over time (Sauer et al., 2010; Palmer et al., 2013). For instance, Deffenbacher and colleagues (2008) found an increase in forgetting as the delay (retention interval) between encoding and testing facial recognition increases. They found this correlation between lengthy retention intervals and decreased face recognition memory also occurs in instances of eyewitness memory.

Additionally, they discussed how factors during the time of encoding (i.e. exposure duration, lighting, witness distance from the perpetrator) may interact with increases in retention intervals to negatively impact a witness's memory strength (Deffenbacher et al., 2008). Essentially, these results suggest that differences in viewing conditions could lead to differences in memory strength as retention interval increases, which can translate to differences in the quality of descriptions provided by witnesses or eyewitness identification accuracy. Thus, this could explain the minor differences found in the number of details provided between the high-quality and low-quality elicitation conditions in this study as the retention interval only consisted of 90 seconds while participants completed the distractor task.

Another factor could be the duration of exposure, even in the low-quality condition, was long enough to allow participants to encode the perpetrator's facial features. In this study, those in the low-quality condition only viewed the perpetrator's face for 10 seconds, which according to much of the eyewitness identification literature is a short exposure duration and would be expected to yield less accurate witness identifications compared to longer exposure durations (Memon et al., 2003; Bornstein et al., 2012). In a study conducted by Memon and colleagues (2003), they found that if a witness viewed the perpetrator's face for only 12 seconds, any identification from a lineup was likely to be unreliable. They also suggested a lineup should not

be conducted when exposure durations were that short unless police had considerable evidence in addition to the witness identification. However, in facial recognition research most exposure durations, or encoding time, range from 3-10 seconds (Carlson et al., 2016). Despite utilising short exposure times of under 10 seconds, various face recognition tests, including the Cambridge Face Memory Test (CFMT), are considered valid measures of face recognition (DeGutis et al., 2023; although see Burns, 2024; Burns et al., 2023) and suggest individuals can encode information regarding previously unfamiliar faces in a matter of seconds (Duchaine & Nakayama, 2006). For example, the findings from Walker-Smith (1978) suggest that while a two-second presentation of a face may not be adequate to allow participants to encode enough features to provide a detailed verbal description, this short exposure time is enough to allow recognition of previously seen facial features and detect any alterations made to those features. Therefore, the ten-second exposure time witnesses had to encode the perpetrator's facial features combined with the short retention interval may have allowed witnesses to recall enough information to provide ample, overall descriptions of the perpetrator.

Though our findings suggest we were successful in manipulating description quality, there are limitations to our study. Firstly, we purposely confounded the description elicitation method and viewing condition to ensure a more pronounced difference in description quality. We did so because the aim of our study was to elicit descriptions that varied in quality for the next study, rather than to test the effects these variables could have on description quality. Additionally, in instances where a default value was missing from either a low- or high-quality description, that missing default value was added to ensure all basic information was present in the description which may have influenced our findings. However, research has shown how omitted default values can negatively impact the quality of a police lineup and, therefore, felt it

important to ensure any missing defaults were included (Lindsay et al., 1994; Demarchi & Py, 2009).

The descriptions differing in quality are important as the perpetrator descriptions created in this study will be used to create MTD lineups in Study 1b. In this next study, two versions of the MTD construction method will be implemented and compared to one another alongside the MTS lineup construction method. As both MTD and MTS methods of filler selection are commonly used to create police lineups, it is important to note how to best elicit detailed person descriptions. It is also important to note that differences in description quality, often seen in the real-world, may impact the construction of MTD lineups. As previous research has cautioned against the use of vague perpetrator descriptions to select fillers, it was important to create descriptions varying in detail to be used during the lineup construction process (Lindsay et al., 1994). By adding this random variable, it will allow for more generalisability of the results from testing the lineups on metrics such as fairness, suspect-filler similarity, and lineup member description-fit.

Study 1b

As previously discussed, the implementation of the MTD method has varied significantly since it was originally proposed (Wells et al., 1993). To our knowledge, Study 1b is the first to test whether blinding the lineup constructor to the suspect's appearance impacts the effectiveness of the MTD method. The descriptions elicited in Study 1a were provided to lineup constructors in both a blinded version of the MTD method (lineup constructors were provided with only the description) and an unblinded version of the MTD method (lineup constructors were provided with the description and an image of the suspect) to create lineups for various perpetrators. Lineup constructors in an MTS condition were provided only with an image of the suspect. We

expected that participants in the unblinded condition would be unable to disregard the suspect's appearance when making their selections, and that their filler selections would, therefore, resemble those made by MTS lineup constructors.

Though the main purpose of this study was to create lineups for use in subsequent studies, we predicted there would little overlap in filler selections between blinded MTD lineups and MTS lineups. We also predicted that the unblinded MTD lineups will have more fillers in common with MTS lineups than with the blinded MTD lineups.

Method 1b

Participants. A sample of 24 Swansea University students participated in this study. Their ages ranged from 18-34 years old, and most were female (18 female (75%), 2 male (8.33%), 2 non-binary (8.33%), 2 preferred not to respond (8.33%)). These individuals were recruited through the online university participant pool and given the choice of course credit or a £10 Amazon voucher in exchange for participation.

Design. This experiment followed a 3 (Lineup construction method: MTS, unblinded MTD, and blinded MTD) x 2 (Description quality: high versus low quality) x 2 (Target presence: target present versus target absent) mixed design. The lineup construction method was manipulated between-subjects, while description quality and target presence were manipulated within-subjects.

Materials

Suspects. The photos of the guilty suspects were used with permission from research conducted by Molinaro et al (2021). These stimuli are openly available via a Creative Commons license (CC-BY 4.0), stimulus materials can be located through Open Science Framework (https://osf.io/dz5pf/?view_only=2b41a6ce5c4d43e0abb08ea1f894d6b2). The photos of innocent

suspects were selected by agreement from CE and RH using the MTD method. This was done in a two-part process. First, CE used the description to create a narrowed pool of 10-12 individuals which best matched all features mentioned in the description for each mock crime video. Next, CE and RH selected the individuals they perceived best matched the perpetrator description for each stimulus set. Any disagreements were discussed, and each individual was compared to the description until agreement was reached, and the innocent suspect selected by consensus.

Face Database. The researchers created a large database of 667 faces that had been pre-sorted into folders by race and then by gender. These images were sourced from multiple face databases, such as the Chicago Face Database, the Face Research Lab London, the FACES database, and the SiblingsDB (Ma, Correll, & Wittenbrink, 2015; DeBruine & Jones, 2017; Ebner, Riediger, & Lindenberger, 2010; Vieira, Bottino, Laurentini, & De Simone, 2014). Where required, permission was obtained by database creators to use images in this research study. All individuals depicted in these databases had given prior consent to their images being used for research purposes. The photographs were frontal facing with neutral expressions. Once these photographs were selected, slight edits were made to ensure they were standardized across all races and genders. These edits included ensuring the individual in the photograph was centred, in front of a white background, and cropped from the neck up. Each image was given an identifier (e.g., “Face 1”) to allow participants to clearly select fillers to include in their lineups.

Procedure. Once participants arrived in the lab, they were greeted and instructed to be seated in front of a computer on the university’s campus. Participants were then given a physical copy of an information sheet to read, followed by an informed consent form. Once participants agreed to participate in the study, they were randomly assigned to a lineup construction condition.

Participants in the blinded MTD condition were ONLY provided with a (high or low quality)

description of the suspect and were provided the following instructions (emphasis in original):

“Please select 5 individuals from the database which **best** match the culprit description below.

Once you have selected the 5 individuals you think **best match the description provided**,

please write down the order in which you would like them presented on the suspect lineup form.

Remember to also assign one spot for the suspect.” Participants who had been assigned to

construct Target-present lineups received the same descriptions as those assigned to create

Target-absent lineups.

Participants in the unblinded MTD condition were provided with a (high or low quality) description of the suspect AND the perpetrator’s photograph (target present condition) or the innocent suspect’s photograph (target absent condition). Participants in this condition were given the same instructions as those in the blinded MTD condition. That is, they were instructed to

“Please select 5 individuals from the database which **best** match the culprit description below.

Once you have selected the 5 individuals you think **best match the description provided**,

please write down the order in which you would like them presented on the suspect lineup form.

Remember to also assign one spot for the suspect.” As with the blinded MTD condition,

participants who had been assigned to construct Target-present lineups received the same

descriptions as those assigned to create Target-absent lineups.

Participants in the MTS condition were ONLY provided with the perpetrator’s photograph (target present condition) or with the innocent suspect’s photograph (target absent condition) and were given the following instructions (emphasis in original): “Please select 5 individuals from the database which **best** match the image of the suspect below. Once you have selected the 5 individuals you think look **most similar to the suspect**, please write down the order in which you would like them presented on the suspect lineup form. Remember to also

assign one spot for the suspect.” Participants in this condition were not presented with a description. However, to enable us to assign participants evenly to all conditions within the factorial design, we yoked participants to the High- and Low-Quality description conditions.

In all conditions, participants were asked to write the identifiers for their five chosen images, or fillers, on a sheet of paper to indicate their positioning of the fillers and suspect within the lineup (*Appendix N*). After each lineup was created, the researcher would confirm with the participant that the identifiers written on the suspect lineup form were those they had intended to select. This was done for several reasons, including making sure the handwriting was legible and ensuring the participant had not incorrectly written down any identifiers. This process continued until the participant had created lineups for each of our four suspects. Upon completion of the fourth and final lineup, participants were debriefed and given a copy of the participant debrief form. The construction of these four lineups took approximately 1 hour to complete.

Results 1b

While the main purpose of this experiment was to create lineups for subsequent experiments, we were also interested in examining self-reported filler selection strategies and to quantify the variability in the number of unique fillers selected when using each of the three lineup constructions methods. To do this, we used the statistical software JASP (v0.18.3, JASP Team (2024)) to perform our analyses. First, we examined the number of unique fillers selected by lineup constructors; this was calculated separately for each stimulus. For context, if every lineup constructor chose the same five fillers, the number of unique fillers for that stimulus set would be five. On the other hand, if every lineup constructor selected a different set of fillers, the number of unique fillers would be 120 (24×5). Thus, the number of unique fillers can serve as an index of variability. Stimulus 1 and stimulus 4 both had just over 30 unique fillers, 33 and 35

respectively, while stimuli 2 and 3 had a higher number of unique fillers used to create the lineups, with 57 and 53 unique fillers selected. These values suggest that the face database used for filler selection contained enough individuals who resembled the suspects (MTS condition), innocent or guilty, or who matched the description of the suspect (MTD conditions), to allow for adequate variation in the lineups created. Therefore, we can be confident that the database was suitable for our research purposes.

In addition to looking at the number of unique fillers selected for the various lineups, we also attempted to quantify the amount of overlap in filler selection between the construction methods. We first counted how many times each filler was selected within each construction method. Using fillers as the unit of analysis, we then correlated these frequencies for each pair of methods. The two MTD methods had the strongest correlation in filler selections, $r(176) = .453$, $p < .001$ suggesting that even though unblinded MTD constructors were provided with the image of the suspect, they also relied on or gave weight to the information provided in the culprit description to make their filler selections. Additionally, we compared the filler selections for the blinded MTD method to the MTS method and found that those methods had the weakest correlation, $r(176) = .167$, $p = .025$, with the correlation for the unblinded MTD method and the MTS method falling in between, $r(176) = .271$, $p < .001$, which suggests the difference in information provided to lineup constructors led to a difference in the fillers selected.

Next, we statistically compared these correlations using the method for comparing correlations from independent samples from Eid, Gollwitzer, and Schmitt (2011). Crucially, filler selections from the unblinded MTD method overlapped more strongly with filler selections from the blinded MTD method than with filler selections from the MTS method, $z = 1.969$, $p = .024$.

Filler Selection Strategies

After lineup constructors finished creating lineups for each of the four stimulus sets, they were asked to describe in detail their filler selection process via the following open-ended question, “please describe the method you used to select the innocent fillers for each lineup you constructed”. This was to give us insight into how lineup constructors select fillers for lineups and how this may have differed based on which construction method was used. While we were most interested in the strategy and process used by those in the unblinded MTD condition, as we wanted to know whether they used both the suspect photo *and* description or truly relied on the description alone, we also wanted to know what features constructors considered most important in the selection process.

The verbatim written responses for each participant in the UMTD condition can be seen in *Table 11*. As seen in the table below, out of the eight lineup constructors using the UMTD method, only one reported basing their selections solely off the descriptions provided. However, there were an additional three lineup constructors who reported using a combination of the suspect image and description to make their selections with one comparing the suspect image to the description to see which traits were in common and then using that as the basis for selection. Another lineup constructor reported using the description to make their initial selections and only then utilised the suspect image to make their final selections. The remaining lineup constructors reported solely using the suspect image to make their filler selections rather than using the description as instructed.

Table 11*Self-Reported Selection Strategies of Unblinded MTD Lineup Constructors*

Information Used	Self-Reported Selection Strategy
Image	"I primarily used the images given and tried to find the closest match. I'm bad at guessing ages so I tried to identify any signs of age in the suspect photo (e.g. wrinkles, grey hair and use that. If the suspect has any distinct characteristics (for example, the White/Hispanic female had some moles on her cheek) I would try and find another image with the same or similar characteristic. Once I was done and I had selected all 5 people to go into the lineup, I would hold up the suspect image sheet to the computer screen and compare to see if they stood out or looked different to the others in any meaningful way." (Participant ID: 3)
Both	"I compared potential fillers that had traits from the description to the image provided. I feel as though this may have been wrong because each question specified to only use the descriptions provided and made no mention of the images, however when so many of the potential fillers share traits the only way to narrow it down further is to use the image of a suspect. When someone had all of the traits from the description and at least vaguely shared a likeness with the image, only then would I choose them as a filler." (Participant ID: 6)
Both	"Initially, on the first set of suspect I selected one at a time mostly compared to the image presented on the description. After compiling images into a lineup, then changed my images as I had noticed I had missed out key information in the description (i.e. one did not have gray hair). After this, I then lined up my suspects and compared them to image and description presented. I mostly searched for likeness between the suspect and the photos. I ensured I selected more than one photo then deducted the ones that did not resemble image/description" (Participant ID: 11)
Image	"I tried to match depending on who looked more like the suspect to be closer to him so as to see if they are identified correctly, and those who look least like the suspect I put them last. This will help identify if the witness can correctly identify the perpetrator." (Participant ID: 12)
Description	"I look for a couple that matches the description given and I rank them based on meets it the most." (Participant ID: 13)
Image	"Looked for similarities to the suspect, such as whether or not they had a beard or long/short hair, as well as ethnicity. Tried to find people that best matched the suspect's general features." (Participant ID: 22)
Both	"I used the descriptors (e.g. white, male) to be in the right category. I then utilised the descriptors on appearance and went through all of the images with this in mind. Ones that matched the description were compared against the suspect's image. The fillers that resembled the suspect more, or the closest I could get to them were selected by me." (Participant ID: 23)
Image	"1) Age based face recognition; 2) hair colour; 3) features (helped to know if on the leaner/heavier side); 4) face characteristics (beard, no beard); 5) looking for components that made them as physically relatable to the suspect." (Participant ID: 24)

In order to determine which features were deemed the most important to match fillers on by lineup constructors, CE and RH independently read and coded the responses provided by participants regarding their selection strategy. We then discussed any disagreements until we reached consensus. The features mentioned most frequently by lineup constructors were hair colour (54.17%), age (37.5%), and race (41.67%). Gender (29.17%) and physical build (29.17%) were also reported by 7 of the 24 lineup constructors as important features used to select fillers. Additionally, features such as face shape, facial hair, hair length, eye colour and shape, and skin texture were also reported, though at lower rates. While not all lineup constructors reported matching fillers to age, race, and gender, the lineups created show that this was taken into account in some capacity in order to prevent the suspect from standing out.

Discussion 1b

The primary purpose of this experiment was to create lineups for subsequent studies using the different filler selection methods. However, we were also interested in examining the variability in the fillers selected by each of the three lineup construction methods, and the filler selection strategies used by each lineup constructor. As expected, there was relatively little overlap in fillers selected for lineups between the blinded MTD and MTS methods. However, contrary to our predictions, fillers selected by lineup constructors using the unblinded MTD method overlapped more with fillers from lineups created using the blinded version of the MTD method than with fillers from MTS lineups. This unexpected finding suggests that the suspect's image may not have exerted a strong influence on filler selections in the unblinded MTD condition. This could be due to the differences in instructions provided to unblinded MTD lineup constructors and those using the MTS method, as unblinded MTD lineup constructors were instructed to select fillers who "best match the culprit description" while MTS lineup

constructors were instructed to select fillers who “look most similar to the suspect”. Though lineup constructors had access to the suspect’s image in both conditions, the instructions may have impacted their decisions filler selections.

However, it is also important to note that most lineup constructors in the unblinded MTD condition reported using the suspect’s image in their filler selection process. This was despite instructions to use the perpetrator description to inform their filler selections. It is perhaps unsurprising, therefore, that there was a moderate correlation between filler selections between the unblinded MTD and MTS conditions. These findings provide some evidence that having access to the suspect’s image may undermine the effectiveness of the MTD method by skewing filler selections toward the suspect’s appearance rather than the perpetrator description.

The findings also show lineup constructors reported that physical features such as hair colour, age, and race were most frequently used to match fillers on, followed by gender and physical build. Other features such as eye shape or colour, hair length or texture, and nose size and shape were rarely reported but may also be important features to match fillers on. However, not all lineup constructors reported matching fillers to age, race, and gender. This finding suggests that default values may also exist in the filler selection process of lineup constructors as these features must have been considered in some capacity to prevent the suspect from standing out from other lineup members. Thus, it is possible some lineup constructors did not mention matching fillers to these features because it seemed obvious as those features are crucial in creating unbiased lineups.

General Discussion

The aim of these studies was to create variable descriptions which could then be used by MTD lineup constructors during the filler selection process. As in real-world scenarios, police do

not always have detailed perpetrator descriptions available. Therefore, we aimed to increase generalisability by providing MTD lineup constructors with both low- and high-quality descriptions. The results of Study 1a show we were successful in eliciting perpetrator descriptions of various quality to be used in the creation of MTD lineups. While the results of Study 1b show we were successful in creating quality lineups using three different filler selection strategies. The number of unique fillers for each stimulus set shows the face database contained enough faces which resembled the suspects (MTS condition) or matched the perpetrator descriptions (blinded and unblinded MTD conditions) to allow for variation in the selected lineup fillers. This is further supported by the results of the filler selection correlations which found the blinded MTD and MTS method had little overlap in the fillers selected which suggests that using different information to select fillers resulted in different lineups.

Additionally, the self-reported selection strategies provided by lineup constructors in the unblinded MTD condition suggest that while there was some overlap in fillers selected between the blinded and unblinded MTD lineups, the lineup constructors were unable to completely disregard the suspect's image, and in doing, so may have impacted their filler selections. This is further supported by the moderate overlap in fillers selected by lineup constructors using the MTS method and those using the unblinded MTD method. These findings taken together offer support for the argument that having access to the suspect's photograph may undermine the effectiveness of the MTD method due to top-down processing (Firestone & Scholl, 2014; Kinchla & Wolfe, 1979). This was previously demonstrated by the example given in the introduction of this chapter (p. 54). The example demonstrated how knowledge of the suspect's appearance may bias lineup constructors in their filler selections by influencing their interpretation of the features described. Thus, it is possible that having access to the suspect's

image during the filler selection process biased the selections made by unblinded MTD lineup constructors. This could be problematic as it could reduce the potential variability among the fillers selected, which research has shown is an advantage of the MTD method (Wells et al., 1993). This is important as a reduction in variability among fillers could negatively impact lineup outcomes, including lineup fairness and witness discriminability.

In conclusion, the findings of this study suggest we were successful in creating stimuli that can be used to investigate the impact blinding MTD lineup constructors has on the effectiveness of the MTD method. The lineups created in this study will be used in three subsequent studies to investigate how filler selection method and the blinding of MTD lineup constructors impacts lineup outcomes. In Study 2, the lineups will be used to investigate how filler selection method and blinding of MTD lineup constructors impacts lineup fairness using the mock witness paradigm. In Study 3a and 3b, the lineups will be evaluated for suspect-filler similarity and lineup member description-fit. Lastly, in Study 4 the lineups will be tested using an eyewitness identification experiment to investigate how filler selection method and the blinding of MTD lineup constructors impacts witness accuracy and confidence, as well as resultant lineup fairness. Each of these future studies will provide insight into which filler selection method yields the fairest lineups, improved lineup outcomes, and whether blinding MTD lineup constructors improves the effectiveness of the method.

Chapter 3

Does filler selection method impact lineup fairness?

Research on eyewitness testimony and suspect lineups has continued to stress the importance of creating and administering a “fair” lineup. But what constitutes a fair lineup? And how does lineup fairness impact eyewitness accuracy? A lineup is considered to be fair if the lineup members resemble the suspect individual to the point that the suspect does not stand out from the others (Wells et al., 2020). However, this is not the only criterion in the creation and administration of a fair lineup. Wells and colleagues (2020) made several recommendations in efforts to improve the quality of suspect lineups and eyewitness identification decisions. These recommendations include providing pre-lineup instructions to witnesses stressing that the perpetrator may not be present in the lineup; a double-blind administration of the lineup in which the lineup administrator does not know who the suspect is; a lineup should have only one suspect alongside a minimum of five appropriate fillers, or known innocent individuals; and a confidence rating should be obtained from the witness regarding their identification decision (Wells et al., 2020). However, it is possible for a lineup to meet each of these criteria and still be biased towards or away from the suspect.

A biased lineup can have serious consequences, especially for innocent suspects, and often results from an issue with the selection of quality fillers. Research has consistently shown that biased lineups (i.e. lineups in which the suspect stands out from other lineup members) result in a higher rate of innocent suspect identifications than fair lineups (Wells, Rydell, & Seelau, 1993; Lindsay et al., 1994). This finding is further supported by research conducted by Carlson, Gronlund, and Clark (2008) which manipulated lineup fairness to investigate how lineup composition and presentation impacts witness identification decisions. Lineup fairness was

manipulated by constructing lineups that varied in fairness: the “biased” lineups were composed of five poor quality fillers with a suspect; “intermediate” lineups were composed of three poor quality fillers and two good quality fillers with a suspect; and “fair” lineups were composed of five good quality fillers with a suspect. All fillers were selected from a large pool of good quality fillers (individuals who matched the perpetrator description) and poor quality fillers (individuals who only partially matched the perpetrator description). The results of this study showed that as lineup fairness decreased, innocent suspect identifications increased. While the fillers in the study by Carlson, Gronlund, and Clark (2008) were all selected by using the MTD method of filler selection, it is important to note that lineups of all levels of fairness can be created using either the MTD or MTS method depending on how each method is employed. This is why it is important to test the lineups for fairness whenever possible.

Testing Lineup Fairness

One commonly used method to measure whether a lineup is fair is by using the mock witness paradigm (Doob & Kirshenbaum, 1973). It has been used in over 250 studies since it was first proposed in the 1970s to present (Lee et al., 2022). In this paradigm, researchers provide a description of a perpetrator and suspect lineup to an individual, a mock witness, who has no knowledge of what the suspect looks like and instructs them to choose the individual from the lineup they think best matches the description provided (Doob & Kirshenbaum, 1973; Mansour et al., 2017). Doob and Kirshenbaum (1973) argued the best way to test for bias in a lineup is to see if a group of “uninvolved people”, mock witnesses, who have never seen the offender before can accurately pick them out among a lineup. If a lineup is fair, then each lineup member should draw a similar number of identifications by the group of mock witnesses. However, if these mock witnesses, who have no knowledge of the suspect, select the suspect at a rate higher than

chance, it suggests that the lineup is biased as the suspect must differ in some way from the other lineup members. These mock witness selections can then be used to give researchers insight into how the lineup might be biased.

Researchers can use the mock witness decisions to calculate a series of measures indicative of the two aspects of lineup fairness, lineup size and lineup bias (Malpass, 1981; Mansour et al., 2017; Lee, Mansour, & Penrod, 2022). Lineup size refers to the number of appropriate fillers and is measured by *effective size* (Malpass, 1981), Tredoux's E (Tredoux, 1998, 1999), and *acceptable lineup members* (Malpass & Lindsay, 1999). Values for measures such as effective size, Tredoux's E, and acceptable lineup members range from 1 to the nominal size of the lineup (Mansour et al., 2017). Effective size measures lineup size by calculating how much the proportion of suspect selections differs from chance, while including only those lineup members selected at least once (Malpass, 1981; Mansour et al., 2017). Tredoux's E is like effective size but also reflects the possibility that lineup members who were not selected could have been chosen if the sample size of mock witnesses was larger (Tredoux, 1998; 1999; Mansour et al., 2017). *Acceptable lineup members* measures the number of lineup members who were selected at a specified rate of chance and is based on the nominal size of the lineup (Malpass & Lindsay, 1999; Mansour et al., 2017). Please see the below table for the formulas for each lineup fairness measure (*Table 12*).

Table 12

Mock Witness Lineup Fairness Measures (adapted from Mansour et al., 2017)

Measure	Formula
Effective Size (Malpass, 1981)	$k_a - \sum_{i=1}^{k_a} \left \frac{n_i - n \frac{1}{k_a}}{2n \frac{1}{k_a}} \right $
Tredoux's E (Tredoux, 1998, 1999)	$\frac{1}{1 - \left(1 - \sum_{i=1}^N \left(\frac{n_i}{n}\right)^2\right)}$
Acceptable Lineup Members (Malpass & Lindsay, 1999)	$\sum_{i=1}^N p_i > \frac{.75}{N}$
Proportion of Suspect Selections (Brigham & Brandt, 1992; Doob & Kirshenbaum, 1973)	$\frac{n_{ss}}{n}$
Functional Size (Wells, Leippe, & Ostrom, 1979)	$\frac{1}{P_{ss}}$
Suspect Bias (Doob & Kirshenbaum, 1973)	$\frac{P_{ss} - \frac{1}{N}}{\sqrt{\frac{1}{N} \left(1 - \frac{1}{N}\right) \frac{1}{n}}}$
Defendant Bias (Malpass, 1981; Malpass & Lindsay, 1999)	$\frac{P_{ss} - \frac{1}{ES}}{\sqrt{\frac{1}{ES} \left(1 - \frac{1}{ES}\right) \frac{1}{n}}}$
Binomial Probability (Tredoux, 1999)	$\left(\frac{n}{n_{ss}}\right) \frac{1^{n_{ss}}}{N} \cdot \frac{(N-1)^{n-n_{ss}}}{N}$

Note: k_a = number of lineup members selected by at least one mock witness, n = number of mock witnesses, n_i = number of mock witnesses who chose that lineup member (i), N = nominal size of lineup, p_i = proportion of selections for lineup member i , n_{ss} = number of suspect selections, P_{ss} = proportion of suspect selections, ES = effective size

Additionally, measures such as the *proportion of suspect selections* (Doob & Kirshenbaum, 1973; Brigham & Brandt, 1992), *functional size* (Wells et al., 1979), *suspect bias* (Doob & Kirshenbaum, 1973), *defendant bias* (Malpass, 1981; Malpass & Lindsay, 1999), as well as *binomial probability* (Tredoux, 1999) are used to measure the bias, or the extent to which the suspect stands out within the lineup (Mansour et al., 2017). While lineup size measures range in

value from 1 to the nominal size of the lineup, lineup bias measures vary greatly from one to the other. The proportion of suspect selections measures bias by taking the number of suspect selections and dividing that number by the number of mock witnesses, while functional size is the inverse of the proportion of suspect selections (Doob & Kirshenbaum, 1973; Wells et al., 1979; Mansour et al., 2017). Suspect bias and defendant bias are both z-scores comparing the proportion of suspect selections to chance using either the inverse of nominal size (suspect bias) or using the inverse of effective size (defendant bias) to indicate lineup bias (Doob & Kirshenbaum, 1973; Malpass, 1981; Mansour et al., 2017). Binomial probability simply refers to the probability of the proportion of suspect selections that occur using binomial distribution (Tredoux, 1999; Mansour et al., 2017).

However, further research on the validity of the mock-witness paradigm found that the paradigm is predictive of eyewitness performance for target-absent lineups but not for target-present lineups (Lee et al., 2022). Lee and colleagues (2022) suggested this may be due to eyewitnesses' memory strength for the perpetrator when the guilty perpetrator is present in the lineup. As a result, researchers often deem resultant fairness to be a better measure of lineup fairness than mock witness-derived estimates. Resultant fairness, or resultant lineup size, is another method used to assess lineup fairness and is done after eyewitness, rather than mock witness, identifications have been made (Quigley-McBride & Wells, 2021). This measure of fairness uses the actual eyewitness identification decisions from the target-absent lineups and examines the distribution of identifications made for the innocent suspect and fillers. These measures provide researchers with information regarding bias towards or away from the innocent suspect and can potentially reveal that a lineup, previously deemed fair during the mock-witness paradigm, was indeed biased toward the innocent suspect. While this measure better predicts

eyewitness performance, it does have a major drawback as it can only be done after the eyewitness identifications have been made and, therefore, cannot be used to measure lineup fairness prior to conducting the identification procedure (Lee, Mansour, & Penrod, 2025). For this reason, resultant lineup fairness will be measured in Chapter 5 and will be compared to the mock witness lineup fairness estimates from this chapter.

Filler Selection Method and Lineup Fairness

Research on lineup fairness has shown that certain methods used to measure lineup fairness, such as the mock-witness paradigm, may be sensitive to variability in lineup construction (Mansour, Beaudry, Kalmet, Bertrand, & Lindsay, 2017). In a study investigating the impact of mock-witness procedures on lineup fairness outcomes, Mansour and colleagues (2017) found that target-presence and type of description used impacted lineup fairness. The results suggested that both lineup size and lineup bias measures indicated target-present lineups as being less fair compared to target-absent lineups. Additionally, they found that lineup size measures indicated lineups were fairer when mock-witnesses read individual descriptions rather than combined, or modal descriptions; but lineup bias measures indicated individual descriptions yielded higher bias ratings. These findings suggest lineup fairness measures are in fact sensitive to the methodologies employed during the mock witness paradigm.

This is further supported by research by Jones (2021), who investigated the effect of filler selection method on lineup fairness through four description-matched methods, comparing them to a suspect-matched method. These four description-matched lineups were based on the type and number of features used to match fillers to the perpetrator description, with GeneralCharacteristics' lineups created matching four general characteristics from the perpetrator description, '4Internal' lineups created matching four internal features (e.g. eye

colour/shape, nose, mouth) from the description, ‘4External’ lineups matching four external features (e.g. hair colour/style, face shape) from the description, and ‘2Internal2External’ lineups matching two internal and two external features from the description (Jones, 2021). The findings of this study suggest that filler selection method does indeed impact lineup fairness estimates, with description-matched lineups yielding significantly higher E values than suspect-matched lineups. It is worth noting that while both effective size and Tredoux’s E are positively associated with suspect-filler similarity, Tredoux’s E does not measure suspect-filler similarity. Instead, it is indicative of how many individuals present in the lineup fit the general characteristics of the suspect, or perpetrator description in the case of the mock-witness paradigm (Fitzgerald, Oriet, & Price, 2015; Jones, 2021). As a result of this, description-matched lineups may yield fairer lineups because all lineup members present in the lineup match the perpetrator description provided to mock-witnesses, whereas this may not be the case for fillers selected based on their similarity to the suspect.

It stands to reason then that lineups created using the MTD method would yield larger lineup effective size measures compared to lineups created using the MTS method. Additionally, as effective lineup size increases, one would expect lineup bias measures to decrease (Mansour et al., 2017). Therefore, if all lineup members in MTD lineups are more likely to match the description provided by mock witnesses, we would expect fairer lineups as demonstrated by larger lineup size measures (effective size and Tredoux’s E) and smaller lineup bias measures (proportion of suspect selections) compared to MTS lineups. With that said, the findings of Jones (2021) and Mansour et al. (2017) highlight how the various lineup fairness measures may be sensitive to differences in methodology, including the methods used to select fillers for the lineups shown to mock witnesses, which we aim to explore further in this study.

Study aims

The purpose of this experiment is to investigate how filler selection method impacts lineup fairness through the use of the mock witness paradigm. Based on research by Jones (2021), we predict that the blinded MTD method will yield larger Tredoux's E and effective size values, and therefore fairer lineups, compared to the MTS method. We also predict that lineup fairness as measured by Tredoux's E and effective size will be similar between the unblinded version of the MTD method and the MTS method as our lineup constructors (Chapter 2) reported using information gained from the suspect's image during the filler selection process.

Method

Participants. A sample of 810 participants were recruited through online methods such as the university participant pool, Prolific, and social media sites such as Facebook. This sample size ensured that each lineup would be viewed by between 100 and 150 witnesses, which was sufficient to calculate stable lineup fairness estimates. This sample size falls within the range commonly seen in mock witness tasks, where lineups are often evaluated by anywhere from 20 to 200 mock witnesses (Carlson et al., 2019; Mansour et al., 2017; Menne et al., 2023).

Participants ranged in age from 18-83 years old ($M = 41.51$ and $SD = 14.33$) with an even spread among gender (females = 51.85%, males = 46.91%, other/prefer not to respond = 1.23%).

University students were compensated with course credits and those recruited from Prolific were paid in accordance with Prolific guidelines, which suggested participants be paid £9.00/hour.

Design. This experiment followed a 3 (Lineup construction method: MTS, Unblinded MTD, Blinded MTD) x 2 (Target Presence: Target Present v Target Absent) within-subjects design. For each target, one of six lineups was selected at random and presented for participants to make a mock witness decision. This was repeated for each of the four targets.

Materials. The materials used in this study were created in previous experiments. The high-quality descriptions of the suspect provided to the mock witnesses were created in Study 1a, while the 6-person lineups provided for the identifications were created in Study 1b. Of the 96 lineups created in experiment 2a, 24 lineups were randomly selected to be evaluated for lineup fairness, allowing for six lineups per target. Evaluating all 96 lineups for fairness would have required an infeasibly large sample, as each lineup needed to be evaluated by 100-150 participants to produce stable lineup fairness estimates.

Procedure. This study was conducted online using the experiment-building platform Gorilla (www.gorilla.sc; Anwyl-Irvine, Massonnié, Flitton, Kirkham & Evershed, 2018). Once participants clicked on the study link, they were redirected to an information screen, followed by an informed consent form. Once participants agreed to participate, they then viewed an instruction screen. The instruction screen advised participants they would be provided with a person description on a screen displayed above images of six faces. They were then instructed to choose the face they thought best matched the person's description, as well as answer the two questions following their selections. After reading the instruction page, participants then viewed one lineup for each of the four suspects. Each lineup was presented simultaneously below the high-quality modal witness description that was used by the lineup constructors in the MTD conditions. For each lineup, participants were asked to select the person they believed best matched the description provided. Once they selected a face, they were then asked a series of decision strategy questions including how easy or difficult they found making an identification, which they rated on a 5-point Likert scale ranging from 'very difficult' to 'very easy' (*Appendix O*). Additionally, participants were presented with five decision strategy statements: "The face I chose just "popped out" at me", "I compared the photos to each other in order to narrow the

choices”, “I first eliminated the ones that didn’t seem to fit, then choose among the rest”, “Upon reading the description, I formed an image of the person in my head. I then selected the person who best matched that description”, “I based the judgment on specific facial features” and instructed to select ‘yes’ or ‘no’ if the statement described their decision-making process (*Appendix N*). These statements came from research by Dunning and Stern (1994) regarding decision processes, though we did adapt the wording slightly on a few of the statements for them to be a better fit for a mock witness task. Upon making the last suspect identification and answering the corresponding questions, participants were debriefed and told they could exit the survey. This study took approximately 5 minutes to complete.

Results

We used the statistical software JASP (v0.18.3, JASP Team (2024)) to perform our statistical analyses. First, we examined the proportion of mock witness choices that fell on the suspect. Collapsed across target-presence, suspects were chosen more frequently in MTS lineups, $M = 0.37$, $SD = 0.48$, than in MTD lineups, whether blinded, $M = 0.20$, $SD = 0.40$, or unblinded, $M = 0.28$, $SD = 0.45$. There was also evidence of considerable variability across stimuli. Both the guilty, $M = 0.36$, $SD = 0.48$, and innocent, $M = 0.74$, $SD = 0.44$, suspects for Stimulus 3 were chosen at a much higher rate compared to the other suspects except for the innocent suspect for Stimulus 4, $M = 0.39$, $SD = 0.49$.

We conducted a mixed effects logistic regression to investigate whether the variables of construction method and target presence influence the probability of a suspect being chosen from a lineup, with lineup construction method and target presence as fixed effects variables and stimulus as a random intercept. These analyses used sum, or effect, contrast coding to allow us to investigate any main or interaction effects (Brehm & Alday, 2022).

We found a main effect of construction method on suspect chosen, $\chi^2 = (2, N = 3224) = 54.21, p < .001$, suggesting the lineup construction method used impacts the likelihood that the suspect will be chosen by a mock witness. Further investigation of this main effect using the “estimated marginal means” function in JASP found the suspect was less likely to be chosen from lineups created using the blinded MTD method ($M = 0.17, SE = 0.06$) compared to lineups created using either the MTS method ($M = 0.32, SE = 0.09$), $\beta = -0.15, SE = .04, z = -3.90, p < .001$, or the unblinded MTD method ($M = 0.23, SE = 0.08$), $\beta = -0.06, SE = .02, z = -2.51, p = .01$. Additionally, the suspect was less likely to be chosen from lineups created using the unblinded MTD method than lineups created using the MTS method, $\beta = -0.09, SE = .03, z = -3.16, p = .003$.

A main effect of target presence on suspect chosen was also found, $\chi^2 = (1, N = 3224) = 254.32, p < .001$. Contrast analysis showed innocent suspects ($M = 0.39, SE = 0.10$) were more likely to be chosen by mock witnesses than guilty perpetrators ($M = 0.13, SE = 0.05$), $\beta = 0.26, SE = .05, z = 4.67, p < .001$. Additionally, we found an interaction between construction method and target presence on the likelihood that the suspect would be chosen, $\chi^2 = (2, N = 3224) = 38.12, p < .001$. These results suggest that the lineup construction method used, as well as whether the target is present, impacts how often the suspect is chosen from the lineup by mock witnesses.

To explore this interaction, we looked at the main effect of construction method separately for target present and target absent lineups. We found that for target absent lineups, construction method significantly impacted the likelihood of the innocent suspect being chosen $\chi^2 = (2, N = 1624) = 118.81, p < .001$. The suspect was less likely to be selected from lineups created using the blinded MTD method ($M = 0.21, SE = 0.09$) compared to lineups created using

either the MTS method ($M = 0.57$, $SE = 0.12$), $\beta = -0.35$, $SE = .05$, $z = -7.18$, $p < .001$, or the unblinded version of the MTD method ($M = 0.40$, $SE = 0.12$), $\beta = -0.19$, $SE = .05$, $z = -4.04$, $p < .001$. Innocent suspects were also less likely to be selected from lineups created using the unblinded MTD method compared to those created using the MTS method, $\beta = -0.16$, $SE = .03$, $z = -4.93$, $p < .001$. When the target was present in the lineup, the effect of construction method was not significant, $\chi^2 = (2, N = 1604) = 2.42$, $p = .30$. It is also worth noting the random intercept for Stimulus accounted for a moderate proportion of the total variance in the probability of a suspect being chosen from a lineup, $\sigma^2 = 0.72$, $SD = 0.848$, $ICC = .42$. In addition to investigating suspect choosing, lineup fairness and the use of decision strategies were also investigated in this study.

Table 13

Mean Proportion and Standard Deviation of Suspect Chosen by Stimulus

	1		2		3		4	
	TP	TA	TP	TA	TP	TA	TP	TA
Proportion Suspect Chosen	0.117	0.212	0.118	0.254	0.364	0.744	0.043	0.390
Std. Deviation	0.322	0.409	0.324	0.436	0.482	0.437	0.202	0.488
Variance	0.104	0.167	0.105	0.190	0.232	0.191	0.041	0.238

Lineup Fairness

The statistical analyses performed to investigate lineup fairness were performed using R Statistical Software (v4.1.2; R Core Team 2021) via the statistical package *r4lineups* (v0.1.1; Tredoux and Naylor, 2018). Lineup fairness is important to investigate, as unfair lineups could increase the likelihood of misidentifications, and can be measured through outcome variables

such as Tredoux's E, effective size, and the proportion at which the suspect was chosen from the lineup by mock witnesses (Tredoux, 1998). Malpass (1981) argued that it is crucial to know how many plausible foils, or known innocent fillers, are included in a lineup to evaluate its fairness. Thus, the measure of effective size was proposed. Effective size attempts to use the distribution of identifications across the lineup to give insight into the overall constitution of the lineup. It does this by taking the sum of the absolute value of the difference between observed and expected values, and dividing the sum by 2 in the calculation reflects the sum of differences where observed values – expected values is greater than 0.

However, this measure assumes that null foils are implausible when in reality null foils, or foils who were not selected by any mock witnesses, have a positive probability of occurring in instances where there is a larger lineup size and there is a smaller number of mock witnesses evaluating the lineup (Tredoux, 1998). Tredoux's E is a modification of Malpass (1981) effective size formula in which has the benefit of a known sampling distribution, where the measure has a maximum, which corresponds to the number of lineup members, and a minimum of 1 as mock witnesses are typically required to make an identification (Malpass, Tredoux, & McQuiston-Surrett, 2007). This means that if some lineup members draw more choices than others, the value of Tredoux's E will move away from the maximum of the sampling distribution and move towards 1. Ideally, all lineup members should have an equal chance at being selected by mock witnesses who have little information to base their decisions on; however, this is not always the case. Lineups which have an effect size or Tredoux's E value closer to the nominal size of the lineup, in other words has a value closer to the total number of members in the lineup, are considered to be fairer as each lineup member was chosen at a similar rate.

The lineup fairness measures of Tredoux's E and effective size were calculated for each

of the 24 lineups tested. These were then aggregated for each type of lineup; the means for these fairness measures are displayed in Table 14. Overall, the blinded MTD method and the unblinded MTD method yielded similar estimates of Tredoux's E. The MTS method, on average, yielded a lower Tredoux's E value suggesting both the MTD conditions yielded fairer lineups than the MTS method. This is further supported as the effective sizes for the three construction methods followed a similar pattern to that seen with Tredoux's E, with both the blinded MTD and unblinded MTD methods yielding similar effective sizes, though the blinded MTD effective size was slightly higher than that of the unblinded MTD method, and the MTS method having the lowest effective size. The table displaying the descriptive statistics for Tredoux's E and effective size for each construction method can be seen below (*Table 14*).

Table 14

Descriptive Statistics for Construction Method, Collapsed Across Target Presence

	Tredoux's E			Effective Size		
	BMTD	MTS	UMTD	BMTD	MTS	UMTD
Mean	3.316	2.789	3.321	3.885	3.432	3.725
95% CI Mean Upper	3.885	3.765	4.249	4.258	4.293	4.395
95% CI Mean Lower	2.748	1.812	2.394	3.512	2.571	3.056
Std. Deviation	0.680	1.168	1.109	0.446	1.030	0.801

Note: BMTD = blinded match-to-description method, UMTD = unblinded match-to-description method, and MTS = match-to-suspect method.

The tables below show lineup fairness estimates for target-present lineups and target-absent lineups separately. For target present lineups (Table 15), both MTD methods, blinded and unblinded, had higher Tredoux's E and effective size values, than the MTS method suggesting that both of the MTD methods yielded fairer lineups when the suspect was present. For both fairness estimates, the lineups created using the blinded MTD method yielded higher values than the MTS method. However, contrary to our predictions, lineups created using the unblinded

MTD method yielded fairness estimates that were more similar to the blinded version of the MTD method than to the MTS method.

Table 15

Descriptive Statistics for Construction Method Across Target Present Lineups

	Tredoux's E			Effective Size		
	BMTD	MTS	UMTD	BMTD	MTS	UMTD
Mean	3.438	2.865	3.279	4.057	3.528	3.755
95% CI Mean Upper	3.750	4.317	4.728	4.484	4.974	4.659
95% CI Mean Lower	3.126	1.413	1.831	3.631	2.083	2.851
Std. Deviation	0.196	0.912	0.910	0.268	0.909	0.568

Note: BMTD = blinded match-to-description method, UMTD = unblinded match-to-description method, and MTS = match-to-suspect method.

Table 16

Descriptive Statistics for Construction Method Across Target Absent Lineups

	Tredoux's E			Effective Size		
	BMTD	MTS	UMTD	BMTD	MTS	UMTD
Mean	3.195	2.712	3.364	3.713	3.336	3.696
95% CI Mean Upper	4.786	5.143	5.635	4.604	5.364	5.419
95% CI Mean Lower	1.603	0.281	1.093	2.823	1.308	1.973
Std. Deviation	1.000	1.528	1.427	0.560	1.274	1.083

Note: BMTD = blinded match-to-description method, UMTD = unblinded match-to-description method, and MTS = match-to-suspect method.

For the target absent lineups (Table 16), a similar pattern is observed as for the target-present lineups; that is, the MTS method yielded lower Tredoux's E values and had a larger confidence interval than both MTD methods. Unlike in target present lineups, the unblinded MTD method yielded a Tredoux's E value that was slightly higher than that of the blinded MTD method suggesting that when the target was absent from the lineup, lineups constructed using the

unblinded MTD method were fairer. However, it is important to note that the unblinded MTD method has a wide confidence interval, similar to that of the MTS method, which suggests that there may be more variability in whether the suspect was chosen by mock witnesses across the various stimuli.

Table 17

Lineup Fairness Values for Each of the 24 Lineups

Lineup	Stimulus	Method	Target Presence	Tredoux's E	Lower CI	Upper CI	Effect Size	Lower CI	Upper CI	Proportion Suspect Chosen	Lower CI	Upper CI
1	1	UMTD	TP	3.493	2.12	4.618	3.958	3.540	4.375	0.0347	0.006	0.069
2	1	MTS	TP	3.972	2.467	4.899	4.523	3.938	4.769	0.169	0.107	0.230
3	1	BMTD	TP	3.170	2.082	3.94	4.411	3.922	4.722	0.162	0.103	0.222
4	1	UMTD	TA	2.362	1.13	3.836	3.1	2.9	4.15	0.375	0.283	0.458
5	1	MTS	TA	3.176	1.892	4.391	3.92	3.399	4.344	0.176	0.112	0.24
6	1	BMTD	TA	2.592	1.601	3.605	3.436	3.014	4.021	0.105	0.056	0.162
7	2	UMTD	TP	4.477	2.574	5.771	4.458	4.053	4.782	0.053	0.022	0.091
8	2	MTS	TP	2.197	1.289	3.193	3.129	2.677	3.758	0.161	0.096	0.225
9	2	BMTD	TP	3.424	1.848	4.361	4.112	3.591	4.536	0.144	0.088	0.2
10	2	UMTD	TA	4.831	3.959	5.704	4.900	4.384	5.218	0.337	0.264	0.410
11	2	MTS	TA	4.628	3.341	5.35	4.790	4.232	5.069	0.341	0.255	0.418
12	2	BMTD	TA	3.774	1.969	4.878	3.945	3.534	4.363	0.089	0.041	0.137
13	3	UMTD	TP	2.635	1.576	4.195	3.402	3.084	3.850	0.495	0.389	0.575
14	3	MTS	TP	3.247	2.207	4.273	3.989	3.551	4.339	0.372	0.284	0.452
15	3	BMTD	TP	3.537	2.036	4.534	3.896	3.470	4.200	0.258	0.180	0.329
16	3	UMTD	TA	1.937	1.131	3.148	2.517	2.345	3.388	0.676	0.581	0.741
17	3	MTS	TA	1.148	1.065	2.129	1.906	1.815	2.996	0.932	0.872	0.962
18	3	BMTD	TA	2.132	1.311	3.015	3.1	2.823	3.6	0.623	0.527	0.692
19	4	UMTD	TP	2.510	1.469	3.649	3.2	2.8	3.711	0.022	0	0.044
20	4	MTS	TP	2.041	1.152	3.763	2.472	2.377	3.614	0.008	0	0.023
21	4	BMTD	TP	3.620	2.623	4.293	3.808	3.367	4.294	0.095	0.044	0.147
22	4	UMTD	TA	4.322	2.563	5.589	4.265	3.897	4.602	0.272	0.203	0.340
23	4	MTS	TA	1.895	1.312	2.69	2.727	2.272	3.409	0.712	0.621	0.779
24	4	BMTD	TA	4.278	2.513	5.71	4.372	4.069	4.848	0.193	0.131	0.263

Note: BMTD = blinded match-to-description condition, MTS = match-to-suspect condition,

UMTD = unblinded match-to-description condition. Confidence intervals (CI) are based on a 95% confidence rate.

Lineup fairness was investigated through statistical analyses on each of the 24 lineups, as each lineup varies by lineup construction method, target presence, and stimulus (Table 17). The table below breaks down each lineup by stimulus, construction method used, and target presence, and provides the outcome variables and confidence intervals associated with them. While several measures can be used to assess lineup size and lineup bias, Tredoux's E, effective size, and the proportion at which the suspect was chosen from the lineup are the most reported measures of lineup fairness within the literature. As such, we focused on these three measures to assess lineup size (i.e. Tredoux's E and effective size) and lineup bias (i.e. proportion suspect chosen). The use of these three measures allowed us to investigate both aspects of lineup fairness, as well as to directly compare findings to existing research.

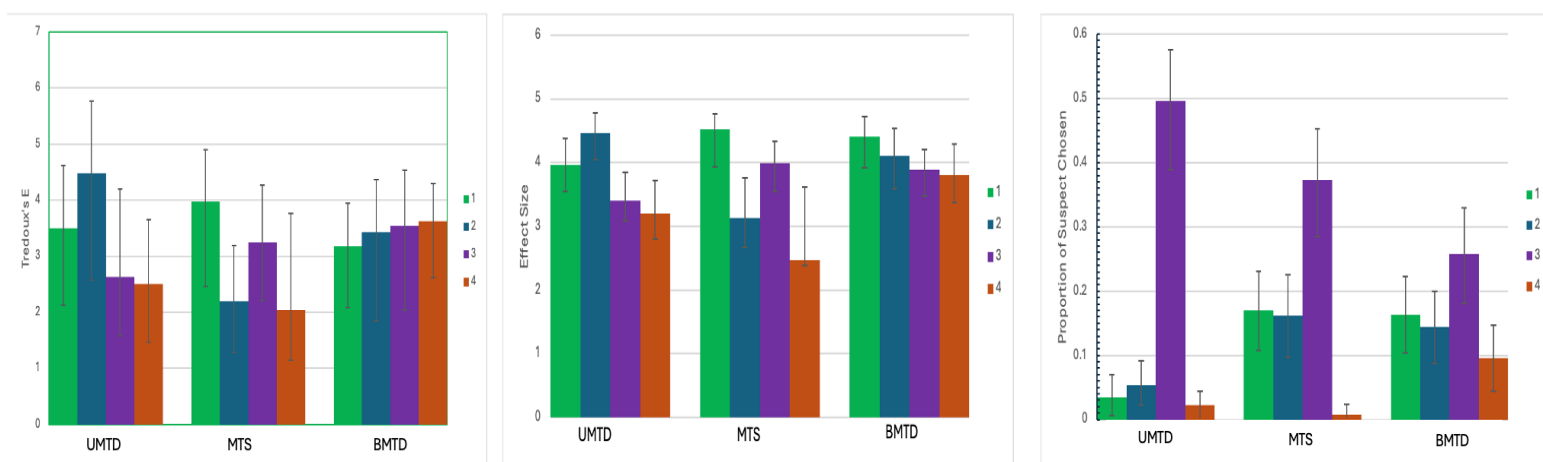
We hypothesised lineups created using the blinded version of the MTD method would yield fairer lineups as demonstrated by larger Tredoux's E and effective size estimates compared to MTS lineups. We also hypothesised lineups created using the unblinded version of the MTD method would perform more similarly to the lineups created using the MTS method than those created using the blinded MTD method. We found for both target present and target absent lineups, the blinded MTD method yielded lineups with larger Tredoux's E and effective size values compared to MTS lineups, supporting our hypothesis. However, contrary to our hypothesis, lineups created using the unblinded version of the MTD method performed more similarly to those created using the blinded version of the MTD method rather than the MTS method.

As seen in the figure below (*Figure 9*), in target-present lineups, the effective size and Tredoux's E' values for the blinded MTD method are more consistent across the various stimuli, while there is greater variability in both the unblinded MTD and MTS methods. However, it

appears the results regarding the proportion of the suspect being chosen from the lineup follow a similar pattern for both blinded MTD and MTS construction methods across the various stimuli (Figure 9). It also appears the likelihood of the suspect being chosen from a unblinded MTD lineup is extremely low, except in the case of Stimulus 3 (Figure 9). However, when the target is absent from the lineup, the three methods appear to follow similar patterns regarding Tredoux's E and effective size (Figure 10). With that said, it is important to note that the innocent suspect is less likely to be chosen from a blinded MTD lineup compared to both unblinded MTD and MTS lineups (Figure 10).

Figure 9

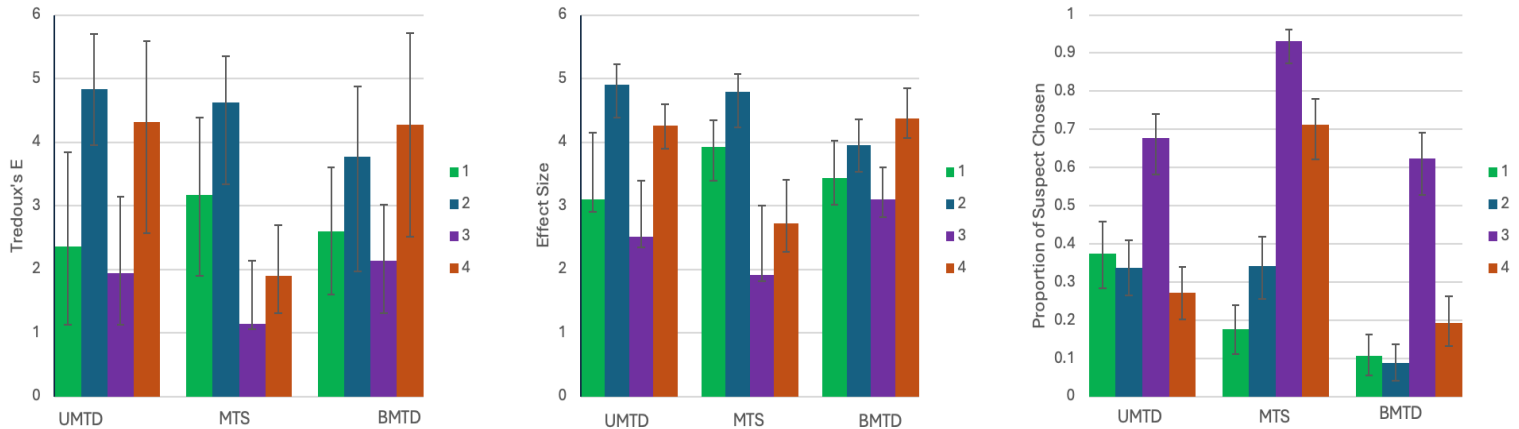
Lineup Fairness Measures by Construction Method and Stimulus for Target Present Lineups



Note: BMTD = blinded match-to-description condition, MTS = match-to-suspect condition, UMTD = unblinded match-to-description condition. From left to right figures depicting line fairness measures: Tredoux's E, effective size, and proportion of suspect chosen.

Figure 10

Lineup Fairness Measures by Construction Method and Stimulus for Target Absent Lineups



Note: BMTD = blinded match-to-description condition, MTS = match-to-suspect condition, UMTD = unblinded match-to-description condition. From left to right figures depicting line fairness measures: Tredoux's E, effective size, and proportion of suspect chosen.

Decision Strategy

In addition to investigating lineup fairness, we conducted an analysis of variance (ANOVA) to investigate whether lineup construction method and target presence impact difficulty of decision, or how hard it was for mock witnesses to make an identification. We found a main effect of lineup construction method on difficulty, $F(2, 1) = 8.36, p < .001$, suggesting lineup construction method impacts how difficult mock witnesses found making an identification. The main effect of target presence on difficulty was also found, $F(1, 2) = 25.16, p < .001$, suggesting mock witnesses found it more difficult to make a decision when the target was present in the lineup than when the target was absent. We also found an interaction between construction method x target presence on identification difficulty, $F(2, 1) = 6.34, p = .002$. These results suggest that difficulty in identifying a suspect from a lineup is impacted by whether the target is absent or present in the lineup, and what lineup construction method was used to create

it.

Post hoc comparisons using the Tukey HSD test were then conducted to further investigate the interaction between construction method and target presence. This analysis showed that mock witnesses found making an identification from MTS lineups ($M = 2.98$, $SE = 0.03$) significantly easier than making one from either the blinded MTD ($M = 2.83$, $SE = 0.03$), $t(3217) = 3.45$, $p = .002$, or unblinded MTD lineups ($M = 2.82$, $SE = 0.03$), $t(3217) = 3.65$, $p < .001$, meanwhile there was no significant difference between the blinded MTD and unblinded MTD lineups, $t(3217) = 0.22$, $p = .97$. Additionally, the post hoc comparisons showed that mock witnesses found making identifications easier from target absent lineups ($M = 2.97$, $SE = 0.02$) than from target present lineups ($M = 2.79$, $SE = 0.02$), $t(3217) = 5.02$, $p < .001$.

Additionally, we conducted a series of independent samples t -tests to investigate whether decision strategies were associated with difficulty of decision (*Table 18*). On average, difficulty in making an identification decision was rated as easier on trials in which the participant endorsed the strategy “The face I chose just ‘popped out’ at me” than on trials in which this strategy was not endorsed, $t(3221) = -20.40$, $p < .001$. The was also true of trials in which the mock witness endorsed using the decision strategy “Upon reading the description, I formed an image of the person in my head. I then selected the person who best matched that description”, $t(3221) = -6.58$, $p < .001$. However, mock witnesses who reported using the decision strategies of “I compared the photos to each other in order to narrow down the choices”, $t(3221) = 4.82$, $p < .001$, and “I first eliminated the ones that didn’t seem to fit, then chose among the rest”, $t(3221) = 3.91$, $p < .001$, found it significantly more difficult to make an identification decision than those who did not endorse those strategies.

Table 18*Mean Proportion and Standard Deviation for Decision Strategy Usage*

Decision Strategy	M	SD
“I compared the photos to each other in order to narrow the choices”	0.82	0.38
“The face I chose just “popped out” at me”	0.22	0.42
“I first eliminated the ones that didn’t seem to fit, then choose among the rest”	0.74	0.44
“Upon reading the description, I formed an image of the person in my head. I then selected the person who best matched that description”	0.27	0.45
“I based the judgment on specific facial features”	0.70	0.46

We also conducted exploratory analyses to investigate whether there were any associations between lineup construction method and decision strategy used. We conducted separate one-way ANOVAs for each of the five decision strategies, with lineup construction method as the independent variable and decision strategy as the dependent variable, applying a Bonferroni correction to reduce the risk Type 1 errors. However, none of the results were significant. For completeness in reporting, the findings for each decision strategy are as follows: “The face I chose just ‘popped out’ at me”, $F(2, 3220) = 1.52, p = .22$; “I compared the photos to each other in order to narrow down the choices”, $F(2, 3220) = 1.11, p = .33$; “I first eliminated the ones that didn’t seem to fit, then chose among the rest”, $F(2, 3220) = 2.22, p = .11$; “I based the judgment on specific facial features”, $F(2, 3220) = 0.75, p = .47$; and “Upon reading the description, I formed an image of the person in my head. I then selected the person who best matched that description”, $F(2, 3220) = 1.71, p = .18$. While exploratory, these findings suggest that lineup construction method does not influence which decision strategy is used by mock witnesses to make an identification.

Discussion

Overall, the findings of this study support our hypothesis that the blinded version of the MTD method yields higher Tredoux's E and effective size estimates, or fairer lineups, compared to the MTS method. However, our second hypothesis regarding the performance of the unblinded version of the MTD method was not supported, as it yielded Tredoux's E and effective size estimates more similar to the blinded version of the method than to the MTS method. This suggests that lineup constructors implementing the MTD method, even when unblinded, were successfully able to select fillers that matched the description, thereby creating fair lineups.

The results of this study fall in line with existing research. Previous research by Jones (2021) has found filler selection method influences lineup fairness estimates. Jones (2021) investigated the effect of filler selection method on lineup fairness through the use of several description-matched methods and compared them to a suspect-matched method and found that filler selection method influenced lineup fairness estimates and that description-matched lineups yielded significantly higher E estimates than suspect-matched lineups. While both effective size and Tredoux's E are positively associated with suspect-filler similarity, these lineup fairness estimates measure how many individuals present in the lineup fit the general characteristics of the suspect, or perpetrator description in the case of the mock-witness paradigm (Fitzgerald, Oriet, & Price, 2015; Jones, 2021). Therefore, it could stand to reason that lineup members in MTD lineups have a better match, or goodness of fit, to the physical characteristics mentioned in the perpetrator description provided to mock witnesses compared to lineup members in MTS lineups. As a result, each lineup member in MTD lineups would be likely to be selected for an identification decision, whereas the lineup members in the MTS lineups may resemble the suspect but could be missing some of those physical characteristics mentioned in the description

and, thus, would not be selected by mock witnesses. This suggests that the mock witness paradigm may only truly measure one facet of lineup fairness – the degree to which lineup members match a description and may not properly measure other facets of lineup fairness, such as perceptual similarity.

The use of perceptual similarity ratings has been used as an alternative method to the mock witness paradigm to examine lineup fairness. As suspect-filler similarity ratings are indicative of how much the fillers resemble the suspect, they relate to lineup fairness estimates as fairness estimates are simply measures of how much the suspect stands out from the fillers in the lineup (Lee et al., 2025). A study by Carlson, Gronlund, and Clark (2008) investigating the impact of lineup composition on eyewitness identification outcomes found that as suspect-filler similarity increased, lineup fairness also increased. Therefore, it stands to reason that measuring perceived similarity between suspects and fillers would give insight into another important aspect of lineup fairness.

In addition to investigating the impact of filler selection method on lineup fairness, we also examined how the filler selection method and target presence may impact the likelihood of the suspect being selected by mock witnesses. We found that the suspect, both innocent and guilty, was more likely to be selected from lineups created using the MTS method in comparison to both versions of the MTD method. This further supports the findings of the lineup fairness analysis, as it suggests that the suspect is more likely to stand out among lineup members in lineups created using the MTS method than from lineups created using the blinded or unblinded versions of the MTD method. However, while the lineup fairness estimates were similar for blinded and unblinded MTD lineups, it is worth noting that the suspect was more likely to be chosen from unblinded MTD lineups than from blinded MTD lineups. This finding suggests that

blinding MTD lineup constructors may impact the likelihood of the suspect being chosen from a lineup, but perhaps not to the point it impacts lineup fairness when tested using the mock witness paradigm. With that said, this could be due to difficulties in being able to statistically and inferentially compare the various lineup measures (Tredoux, 1999). As lineup fairness measures tend to be reported as descriptive point estimates without any inferential considerations, it can lead to issues with interpreting and comparing lineup fairness results. However, because common lineup measures (e.g. effective size) are often derived from the mock witness paradigm which relies on an implicit probability model, it can cause traditional inference statistics to be intractable (Tredoux, 1999). Additionally, these measures evaluate different aspects of lineup fairness and are calculated differently which further complicates comparison and interpretation of findings. Research has shown that while many lineup size measures (i.e. Tredoux's E, effective size) often correlate strongly with one another (i.e. convergent validity), lineup bias measures (i.e. proportion of suspect selections, suspect bias, defendant bias) do not (Mansour et al., 2017; Lee et al., 2025). This further highlights the issue, since measures which evaluate the same aspect of lineup fairness, such as lineup bias, may yield contradictory information. Thus, further research is needed to improve methods to compare lineup measures as these metrics continue to be an important source of information within eyewitness research and the criminal justice system.

An analysis of the decision strategy data found that witnesses found it significantly easier to make an identification from lineups created using the MTS method than from lineups created using either version of the MTD method. This finding, along with the proportion of suspect selections, provides further evidence that lineups created using either version of the MTD method yield fairer lineups when measured using the mock witness paradigm. Further research is

needed to determine how other facets of lineup fairness, such as perceived suspect-filler similarity, is impacted by filler selection methods.

In conclusion, the results of this study suggest that filler selection method does impact lineup fairness. Specifically, the MTD method yields fairer lineups as measured by Tredoux's E and effective size than the MTS method. However, blinding the lineup constructors who were implementing the MTD method had little impact on lineup fairness estimates. These findings give us important insight into how filler selection method may impact the overall quality of lineups. Additionally, these findings have important implication as there are currently no set procedures for suspect lineup construction in the United Kingdom, other than ensuring the suspect does not stand out among other lineup members. Therefore, these findings may have the potential to inform and shape police policies and procedures regarding the construction of suspect lineups.

Future Directions

We aim to further investigate how filler selection method influences lineup fairness and overall quality of lineups, as well as witness identification accuracy and confidence. In our next study, we will be investigating how filler selection method impacts suspect-filler similarity, as well as description fit. The results of the following study will give additional insight into how filler selection method may influence other facets of lineup fairness not measured by the mock witness task, as well as how suspect-filler similarity and description fit may impact witness discriminability.

Chapter 4

Suspect-Filler Similarity and Lineup Member Description-Fit

Decades of research on eyewitness memory and suspect lineups have shown the important role lineups play in reducing eyewitness misidentifications. Suspect lineups help reduce the identification of innocent suspects by testing a witness's recognition memory rather than simple recall (Luus & Wells, 1991). This is because, at their core, suspect lineups are a recognition task as they test a witness's ability to discriminate between unfamiliar individuals, or fillers, and the guilty perpetrator. However, this test of recognition only works if the suspected individual, guilty or innocent, is not obviously different from the other lineup members, highlighting the importance of selecting quality fillers. Fillers are considered to be quality if they match features mentioned in the perpetrator description and resemble the suspect enough to keep them from standing out (Wells et al., 2020).

While best practice recommendations by Wells and colleagues (2020) state that the lineup should be composed of “appropriate”, or quality, fillers to prevent the suspect from standing out from the other lineup members, they also caution against selecting fillers who look too similar to the suspect as it may make it too difficult for witnesses to confidently and accurately identify the perpetrator. The issue of lineup member similarity is one that has had much attention in recent years. In a meta-analysis, Fitzgerald et al. (2013) found no significant differences in perpetrator identifications between moderate and high similarity lineups when the guilty perpetrator was present in the lineup. However, when the target was absent from the lineup, the innocent suspect was more likely to be chosen from a moderately similar lineup than from a highly similar lineup. This finding suggests that highly similar lineups may reduce the likelihood of witness misidentifications, while having no negative impact on correct

identifications. However, in a later study Fitzgerald and colleagues (2015) found that when comparing identification rates from moderately high similarity and very high similarity lineups, very high similarity lineups led to a reduction in correct identifications and did not lead to a significant reduction in innocent suspect identifications. This finding suggests that perhaps there is such a thing as too similar when it comes to lineup member similarity.

Lineup Member Similarity

Similarity can be manipulated both directly, by using subjective similarity ratings and morphing technology, and indirectly, through filler selection strategies. Research directly manipulating subjective similarity ratings has found mixed results regarding the impact of lineup member similarity on eyewitness identification accuracy. Brewer and Wells (2006) manipulated suspect-filler similarity through the use of pairwise similarity ratings. In this study, participants witnessed a mock staged crime video in which a thief stole a credit card while a waiter was distracted. They then used a two-lineup design (i.e. each participant saw two separate lineups, one for the thief and another for the waiter) to investigate the impact of suspect-filler similarity on identification outcomes. For the high-similarity lineups for both targets, the thief and the waiter, the seven fillers who were rated as most similar to the suspect by 30 independent observers were selected. For the low-similarity lineups in the thief condition the three fillers from the high similarity lineup that were rated highest in suspect-filler similarity were replaced by fillers with lower similarity ratings. The process for creating the low-similarity lineups in the waiter condition differed slightly due to a lack of existing data on the distribution of filler choices and instead two fillers from the high-similarity lineup were replaced rather than three as in the thief condition.

For target-present lineups, Brewer and Wells' (2006) findings were mixed. For the thief

lineups, they found the correct identification rate was higher for high suspect-filler similarity lineups compared to lower similarity lineups. The opposite was found for the waiter lineups, with the lower similarity lineups yielding higher correct identification rates. However, for target absent lineups, the results showed a higher rate of innocent suspect identifications in the high suspect-filler similarity lineups than in the low similarity lineups for the waiter lineups. There was no difference in innocent suspect identification rates between the two similarity conditions for the thief lineups. These mixed results suggests that the direct manipulation of similarity ratings appears to have a minimal effect on witness identification accuracy.

Sauer, Brewer, and Weber (2008) also manipulated lineup member similarity through the use of pairwise similarity ratings. However, their methods in constructing low and high similarity lineups differed slightly from the methodology used by Brewer and Wells (2006). The low and high similarity lineups created in this study was done by ensuring there was a significant difference in mean similarity ratings between the two similarity conditions for both targets. This was done by focusing on the similarity ratings for the lineup member with the highest similarity rating in the low similarity lineup and making sure it was significantly lower than the lineup member with the lowest similarity rating in the high similarity lineup (Sauer et al., 2008). For both the female and male targets, the high similarity lineups ($M = 3.31$, $SD = 0.27$ and $M = 2.86$, $SD = 0.35$, respectively) were significantly higher in suspect-filler similarity than the low similarity lineups ($M = 1.62$, $SD = 0.12$ and $M = 1.70$, $SD = 0.08$, respectively). They found that lineups with higher lineup member similarity led to a small but significant decrease in correct identifications in target present lineups but had no effect on witness accuracy when the target was absent from the lineup.

As with the previous two studies, Fitzgerald and colleagues (2014) also manipulated

lineup member similarity through the use of pairwise similarity ratings. This study used a methodology similar to that of Sauer and colleagues (2008) to create low and high similarity lineups to investigate the impact of lineup member similarity on identification decisions made by adult and child witnesses. They created two lower-similarity lineups, comprised of fillers with lower suspect-filler similarity ratings ($M = 1.78$ and 1.93 , $SD = 0.35$ and 0.11), and two higher-similarity lineups, comprised of fillers with higher suspect-filler similarity ratings ($M = 4.44$ and 4.76 , $SD = 0.76$ and 0.61). They found lineup member similarity had little effect on correct identifications for adult witnesses. However, high similarity lineups reduced the likelihood of innocent suspect identifications compared to low similarity lineups, suggesting that highly similar fillers may be beneficial when the suspect is innocent.

Additionally, Fitzgerald and colleagues (2014) found that while lineup member similarity had little impact on adult witness accuracy, child witnesses (age 6-12) were very sensitive to differences in lineup member similarity. Specifically, children were significantly better at identifying the guilty perpetrator from low similarity lineups than from high similarity lineups. When the target was absent from the lineup, both child and adult witnesses were more likely to make an innocent suspect identification from low-similarity lineups compared to high-similarity lineups. These findings suggest that increased lineup member similarity may negatively impact correct identifications but also may provide protection against misidentifications of innocent suspects. With that said, it is worth noting that for each of the studies mentioned, the subjective similarity ratings for the higher similarity lineups were around the midpoint of the similarity rating scale rather than at the upper limits, with the lower similarity lineups being at the lower limits of the scale. Therefore, the 'high' similarity lineups may not include lineups to the degree of high similarity that has been cautioned against at least regarding adult witnesses.

Another method to directly manipulate lineup member similarity is the use of morphing software. Fitzgerald et al. (2015) used morphing software to create lineups with very high and moderately high suspect-filler similarity to investigate how the upper bounds of similarity may impact identification decisions. The fillers were created by selecting faces rated as being highly similar to the guilty or innocent suspect and then morphing those faces with that of the guilty or innocent suspect. The moderately high similarity lineups were created by morphing the faces of the fillers with that of the culprit to create new faces that were 60% filler and 40% suspect, while the very high similarity lineups were created by morphing the fillers twice. First, filler faces were morphed with the suspect face to create new faces that were 50% filler and 50% suspect. To avoid the fillers appearing too homogeneous, these faces were then morphed a second time with a different person who had been rated as highly similar to the suspect. Fitzgerald et al. (2015) found that participants were more accurate in identifying the culprit from the moderately high similarity lineup than from the very high similarity lineup; furthermore, the correct identification rate for the very high similarity lineup was very low. The results of this study suggest there is a point at which suspect-filler similarity becomes too high and negatively impacts witness accuracy, however it is not yet clear exactly where that tipping point may be.

Filler Selection Method and Lineup Member Similarity

In addition to direct manipulations of suspect-filler similarity, research has shown filler selection methods can indirectly influence similarity among lineup members. Luus and Wells (1991) argued that selecting fillers based on their physical similarity to the suspect could potentially lead to constructors creating lineups too high in lineup member similarity, making it too difficult for witnesses to make an identification. They stated that the MTS method has no clear cut-off as to how many features to match and when similarity to the suspect is taken too far

it could result in a lineup of clones, in which case it would make distinguishing between guilty and innocent individuals impossible. In efforts to combat this issue with high suspect similarity, Wells et al. (1993) proposed the MTD method of filler selection. They argued that by selecting fillers based on how well they match the description of the perpetrator not only provides a natural cut-off point for matching features but also allows for some variety among the fillers. In theory, this added variation in fillers should improve witness accuracy while still ensuring moderate suspect-filler similarity so that the lineup remains unbiased towards the suspect.

In research comparing the impact of different filler selection methods on lineup outcomes, Tunnicliff & Clark (2000) found that lineups constructed using the MTS method led to higher lineup similarity ratings for both target present and target absent lineups compared to lineups created using the MTD method. For the purposes of this study, lineup constructors consisted of detectives from the local sheriff's department who had at least three years of experience in constructing lineups. These detectives were provided with either a description of the culprit or a photograph of the suspect, depending on which condition they were in, and selected five fillers based on the information provided. Participant witnesses returned two weeks after the staged crime to view a lineup. Additionally, a second group of participants who had not witnessed the staged crime rated the degree of similarity between the suspect and the fillers. This study found that filler selection method impacts lineup member similarity, with MTS lineups yielding higher suspect-filler similarity ratings compared to MTD lineups. However, they did not find any significant differences in correct and false identifications between MTS and MTD lineups.

Darling, Valentine, and Memon (2008) also found lineup member similarity differences between MTS and MTD lineups, with MTS lineups yielding higher suspect-filler similarity

ratings for both guilty and innocent suspects compared to MTD lineups. For the purposes of this study, MTS lineups were created by researchers selecting eight fillers who closely resembled the suspect. MTD lineups, on the other hand, were created by researchers selecting fillers who not only matched the individual witness's description of the perpetrator but also varied as much as possible without violating the description. Participants witnessed an in-person, staged crime and then returned after between 7 and 25 days to make an identification from a lineup. Additionally, two groups of independent raters assessed suspect-filler face pairings for pairwise similarity or description-face pairings for the degree of match between the witness's description and the lineup members (Darling, Valentine, & Memon, 2008). The results of this study support those of Tunnicliff and Clark (2000), with fillers in MTS lineups rated as being more similar to the suspect than fillers from MTD lineups. They also found no significant differences in correct and false identifications between the two filler selection methods. Darling and colleagues (2008) offered a potential explanation for this finding, stating this could be because the similarity ratings for the MTS lineups were not to the degree of similarity which could impair witness accuracy.

Additionally, a review of the literature provides further evidence that suspect-filler similarity is impacted indirectly by manipulating filler selection strategies (Fitzgerald et al., 2015). This review showed that in studies directly comparing the MTS method of filler selection to the MTD method, lineups created using the MTS method consistently yield higher suspect-filler similarity ratings compared to lineups created using the MTD method. In fact, Fitzgerald et al. (2015) went as far as to say that differences in similarity ratings can be inferred by “substituting “moderate similarity” with description-matched and “high similarity” with appearance-matched” when discussing the existing literature on filler selection methods and the impact it has on suspect-filler similarity (Fitzgerald et al., 2015, p. 63). Thus, while the findings

of both studies and the literature review suggest that filler selection method does impact lineup member similarity; they also suggest that as long as lineup member similarity is not pushed to the extreme, it may not negatively impact witness identifications. In other words, while filler selection method may lead to differences in suspect-filler similarity, these differences may not necessarily translate into a difference in lineup outcomes.

Impacts of Suspect-Filler Similarity on Identifications

While lineups with high suspect-filler similarity have been shown to reduce innocent suspect identifications, research has shown that too high suspect-filler similarity may also reduce correct identifications (Fitzgerald et al., 2015). In an article discussing the importance of lineup fillers and how those fillers should be selected, Luus & Wells (1991) argued that there was an “optimal” level of similarity among lineup members. On one hand, low suspect-filler similarity could lead to biased lineups resulting in the suspect standing out from the fillers included in the lineup and potentially increasing innocent suspect identifications. On the other hand, too high suspect-filler similarity could lead to a lineup of clones. This would make it nearly impossible for witnesses to confidently and accurately distinguish between innocent and guilty individuals and result in a reduction of correct identifications. This “optimal” level of similarity would fall somewhere in between these two extremes, thereby ensuring that lineup members are similar enough to prevent the suspect from standing out in the lineup but also have some variation in features in order to allow witnesses to distinguish between the perpetrator and innocent individuals.

As Fitzgerald et al. (2015) demonstrated, it is possible to create lineups with too high suspect-filler similarity to the point it negatively impacts witness accuracy. While this was done through morphing software, it is possible that lineup constructors utilising the MTS method may

end up with highly similar fillers, especially as there is no natural cut-off point as to how many features constructors should match the suspect and fillers on. Wells and colleagues (1993) argued that as the MTS method of filler selection bases filler selection on physical similarity to the suspect, this method may be problematic as it may result in a lineup of individuals who too closely resemble one another and make it too difficult for witnesses to make an identification decision.

While too-high similarity may be problematic, decades of research have shown the consequences of having lineups that are biased or are low in suspect-filler similarity, especially when the suspect is innocent. Research by Wells and colleagues (1993) found that biased lineups led to a higher rate of innocent suspect identifications than fair lineups. This finding was supported by research conducted by Lindsay et al. (1994), which found that biased MTD lineups led to a higher rate of innocent suspect identifications compared to fair MTD and MTS lineups. Additionally, research by Charman, Wells, & Joy (2011) found lineups containing highly dissimilar fillers or “duds” can increase a witness’s confidence in a misidentification. These findings show how low similarity lineups can lead to increased misidentifications and may contribute to wrongful convictions. Therefore, it is important lineup constructors create fair lineups, with an “optimal” level of suspect-filler similarity to ensure witnesses are confident and accurate in their identification decisions.

Lineup Member Description-Fit

While much research has been conducted on the impact of lineup member similarity on witness identifications and lineup fairness, little research has been done on how well lineup members match perpetrator descriptions and how that impacts identification decisions. Research by Darling et al. (2008) is one of the few studies which has collected lineup member description-

fit ratings in addition to suspect-filler similarity ratings. The results of this study found that while suspect-filler similarity was higher in lineups created using the MTS method than in lineups created using the MTD method, lineup members in MTD lineups were rated as matching the description better than lineup members in MTS lineups. However, Darling et al. (2008) did not report whether MTD lineup constructors were blinded to the suspect's appearance when constructing the lineups. If MTD lineup constructors were not blinded to the suspect's appearance, it could have biased their selections of fillers and reduced the effect size of filler selection method on description-fit ratings found in this study (Darling, Valentine, & Memon, 2008). While the researchers provided a potential explanation as to why no significant differences in correct and false identifications between the two filler selection methods were found regarding suspect-filler similarity, no such explanation was provided regarding lineup member description fit (Darling, Valentine, & Memon, 2008). With that said, the results of this study suggest that while filler selection method does impact lineup member similarity and description-fit ratings, it may not be to the degree that it negatively impacts witness accuracy.

However, as discussed in Chapter 3, lineup member fit to the description may impact lineup fairness estimates when utilising the mock witness paradigm to evaluate lineup fairness. The findings of that chapter suggest that the mock witness paradigm may only truly measure one facet of lineup fairness, the degree to which lineup members match a description, and may not properly measure other aspects of lineup fairness, such as suspect-filler similarity. While we found that MTD lineups yielded fairer lineups compared to MTS lineups, as demonstrated by larger Tredoux's E and effective size estimates, we found no difference in lineup fairness estimates between the blinded and unblinded MTD lineups. As the mock witness paradigm may not measure all aspects of lineup fairness, this study uses perceptual similarity ratings and lineup

member description-fit ratings to further investigate how filler selection method may impact lineup fairness. As suspect-filler similarity ratings are indicative of how much the fillers resemble the suspect, they relate to lineup fairness estimates, as fairness estimates are simply measures of how much the suspect stands out from the fillers in the lineup (Lee et al., 2025). Therefore, it stands to reason that measuring perceived similarity between suspects and fillers would give insight into another important aspect of lineup fairness. Additionally, measuring lineup member description-fit may provide further evidence that the mock witness paradigm measures only how many members in the lineup match the description.

Study Aims

The purpose of this experiment is to investigate how filler selection method impacts filler-suspect similarity and description-fit among lineup members. This was done in two studies by having participants rate suspect-filler face pairings for similarity (3a) and rate lineup member photos for description-fit (3b). The suspect and filler photos used in both parts of the study came from the lineups created using the MTS method, the blinded version of the MTD method, and the unblinded version of the MTD method in Experiment 1b.

For Study 3a, as MTS lineup constructors selected fillers who physically resembled the suspect, we hypothesized that the MTS method would yield lineups with higher similarity ratings than lineups created using the MTD method (Hypothesis 1). This hypothesis is in line with the existing literature on lineup member similarity. Additionally, we hypothesised that the unblinded version of the MTD method would yield suspect-filler similarity more similar to the MTS method than to the blinded MTD method (Hypothesis 2). This hypothesis is based on research on top-down processing, which would suggest lineup constructors in the unblinded version of the MTD method would be unable to disregard the suspect's image when selecting fillers.

For Study 3b, we hypothesised that the blinded version of the MTD method would yield higher description-fit ratings among lineup than the MTS method (Hypothesis 3). We also hypothesised that the unblinded version of the MTD method would once again perform more similarly to the MTS method than to the blinded version of the method (Hypothesis 4).

Method 3a

Participants. A sample of 105 participants were recruited through online methods, with 60% of participants recruited through the university participant pool and 40% of them recruited through Prolific. Each participant viewed 90 face pairings; this ensured each pairing was rated by 20 to 30 participants. Power analysis using G*Power v3.1 (Faul et al., 2007) showed this was sufficient to detect a small effect size of $\eta^2 = 0.02$. Participants ranged in age from 18-65 years old ($M = 27.58$ and $SD = 11.41$) with the majority of participants being female (females = 74.29%, males = 22.86%, other/prefer not to respond = 2.86%). University students were compensated with course credits and those recruited from Prolific were paid in accordance with Prolific guidelines, based on the study timing this resulted in £11.15/hour.

Design. This experiment followed a 3 (Lineup construction method: MTS, unblinded MTD, blinded MTD) x 2 (Target presence: target present versus target absent) within-subjects design. Each face pairing shown was a filler with either the perpetrator (target present) or an innocent suspect (target absent). These pairings came from lineups created by the three lineup construction methods in Experiment 1b.

Materials.

Face pairs. The materials used in this study were created in previous experiments. The 358 face pairings came from lineups created in Study 1b, with each face pairing consisting of one filler and one suspect, guilty or innocent. The 358 face pairings were then split into four separate lists

of 89 or 90 via a list randomiser. Participants were randomly assigned one of the four lists, with the face pairings shown in a random order through use of the randomisation function on Gorilla. *PI-20*. The PI-20 is a 20-item self-report questionnaire that is used to distinguish those with prosopagnosia, also known as face blindness, from the general population. It is composed of 20 statements relating to facial recognition experiences, such as “My face recognition ability is worse than most people” and “I often mistake people I have met before for strangers”. These statements are scored on a 5-point scale ranging from 1 (strongly disagree) to 5 (strongly agree), this questionnaire can be found in *Appendix P*.

Procedure. This study was conducted online using the experiment-building platform Gorilla (www.gorilla.sc; Anwyl-Irvine, Massonnié, Flitton, Kirkham & Evershed, 2020). Once participants clicked on the study link, they were redirected to an information screen, followed by an informed consent form. Once participants agreed to participate, they then viewed an instruction screen. The instruction screen advised participants they would be shown a series of face pairs and asked to rate each face pair for similarity on a 7-point scale, ranging from 1 (very dissimilar) to 7 (very similar). They were told to take their time while comparing the faces and that there was no right or wrong answer. They were to repeat this rating process for 90 pairings, with a short break midway through. The face pairings were displayed side by side with the similarity scale displayed below. These pairings remained on the screen until participants submitted their responses. Participants were shown the face pairings in a randomized order. The suspect’s position (left or right) was also randomized. Upon rating the last of the face pairings, participants were asked to complete the PI-20. After completing the PI-20, participants were debriefed and told they could exit the survey. This study took approximately 15 minutes to complete.

Results 3a

We used the statistical software JASP (v0.19.3, JASP Team (2024)) to perform our statistical analyses. As research has shown individuals with prosopagnosia struggle with judging faces for perceived similarity and face matching, we excluded any individuals who self-reported being diagnosed with prosopagnosia and those who scored a 65 or higher on the Prosopagnosia Index (Stantić et al., 2022). The Prosopagnosia Index, or “PI-20”, was used to screen for participants who had either been diagnosed with prosopagnosia or scored 65 or above on the questionnaire – which signifies impairment with face recognition (Shah et al., 2015). Out of the 105 participants in this study, 10 were excluded for scoring 65 or higher on the PI-20 questionnaire. For this analysis, face pairings formed the unit of analysis. Because we needed to assign each face pairing to a single construction method (MTS, blinded MTD or unblinded MTD), we excluded any fillers that were selected by lineup constructors across more than one construction method. For example, if a filler was chosen by both an MTS and a blinded MTD lineup constructor, that filler was excluded from this analysis. For this reason, 192 face pairings were excluded from this analysis.

First, we examined the similarity ratings for each face pairing. As fillers were selected based on their physical similarity to the suspect in the MTS condition, we hypothesised that filler-suspect face pairings coming from MTS lineups would be rated as more similar than filler-suspect face pairings from blinded MTD lineups. Additionally, we hypothesised that filler-suspect face pairings coming from lineups created by the unblinded MTD method would be rated as more similar to one another compared to the blinded MTD method, as lineup constructors had access to the suspect’s image when making their filler selections. Table 19 depicts the mean filler-suspect similarity ratings by lineup construction method, and as predicted, the MTS method

led to higher filler-suspect similarity ratings, in both target present and target absent lineups, compared to the blinded MTD method. However, the mean filler-suspect similarity ratings were only higher in the unblinded MTD condition compared to the blinded MTD method when the target was present in the lineup, only partially supporting our hypothesis.

Table 19

Mean Filler-Suspect Similarity Ratings by Lineup Construction Method

	MTS		BMTD		UMTD	
	TP	TA	TP	TA	TP	TA
Filler-Suspect Similarity	2.842	3.637	2.666	3.266	2.852	3.197
Standard Deviation	0.908	1.209	1.041	1.168	1.077	1.232

We conducted repeated measures analysis of variance (ANOVA) to investigate whether the variables of construction method and target presence influence the similarity ratings, with lineup construction method and target presence as the main effects and interaction term and similarity rating as the dependent variable. We found a significant main effect of Lineup Construction Method on similarity rating, $F(2, 185.96) = 14.45, p < .001$, with a medium effect size, $\eta^2 = 0.04$, suggesting that the lineup construction method used does influence how similarly face pairings are rated. Post hoc test comparisons using the Holm correction showed that the MTS condition ($M = 3.24, SE = 0.10$) yielded significantly higher filler-suspect similarity ratings than either the blinded MTD ($M = 2.97, SE = .10$), $t(93) = 5.09, p < .001$, or unblinded MTD conditions ($M = 3.02, SE = 0.11$), $t(93) = 3.78, p < .001$, supporting our hypothesis that the MTS method would lead to lineups with lineup members more similar to one another than MTD lineups. However, contrary to our hypothesis there was no difference in filler-suspect similarity

ratings between the two MTD methods, $t(93) = 1.18, p = .24$. To investigate this finding further, an exploratory Bayesian 2 (Lineup Construction Method: blinded MTD, unblinded MTD) $\times$ 2 (Target Presence: target present, target absent) repeated measures ANOVA was conducted using JASP (v0.19.3, JASP Team (2024)). This analysis found support for the null hypothesis over the alternative, $BF_{01} = 6.37$, indicating no difference in suspect-filler similarity between the two MTD methods. We also found a significant main effect of Target Presence on similarity rating, $F(1, 93) = 87.99, p < 0.001$, with a large effect size, $\eta^2 = 0.23$. Fillers were rated more similar to the innocent suspects ($M = 3.37, SE = 0.11$) than they were to guilty perpetrators ($M = 2.79, SE = 0.10$).

In addition to both significant main effects, we found an interaction between Lineup Construction and Target Presence, $F(1.81, 168.18) = 9.72, p < .001$, with a small effect size, $\eta^2 = 0.02$ (*Figure 11*). To analyse the interaction further, two one-way ANOVAs were conducted on target present and target absent lineups. Lineup construction method had a significant impact on perpetrator-filler similarity, $F(2, 186) = 4.74, p = .01$. Post hoc test comparisons using the Holm correction showed that the blinded MTD method ($M = 2.67, SE = .11$) yielded significantly lower perpetrator-filler similarity ratings than both the MTS method ($M = 2.84, SE = 0.09$), $t(93) = -2.40, p = .04$, and the unblinded MTD method ($M = 2.85, SE = 0.11$), $t(93) = -2.91, p = .01$. There were no significant differences in perpetrator-filler similarity between the MTS and unblinded MTD conditions, $t(93) = -0.15, p = .88$.

Lineup construction method also had a significant impact on innocent suspect-filler similarity ratings, $F(2, 186) = 17.68, p < .001$. Post hoc test comparisons using the Holm correction showed that the MTS condition ($M = 3.64, SE = 0.12$) yielded significantly higher innocent suspect-filler similarity ratings than either the blinded MTD ($M = 3.27, SE = .12$), $t(93)$

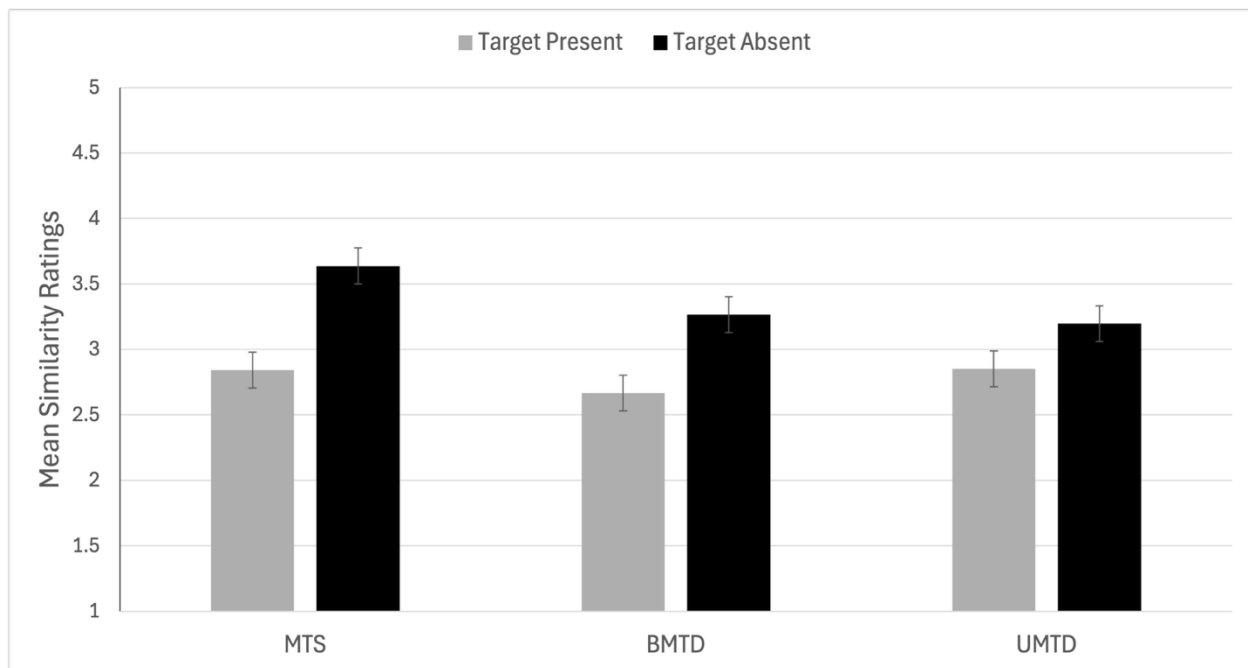
= 4.34, $p < .001$, or unblinded MTD conditions ($M = 3.20$, $SE = 0.13$), $t(93) = 5.81$, $p < .001$.

There were no significant differences in innocent suspect-filler similarity between the blinded and unblinded MTD conditions, $t(93) = 0.89$, $p = .38$. To investigate this finding further, an exploratory Bayesian paired samples T-Test was conducted comparing the blinded MTD and unblinded MTD methods for target absent lineups using JASP (v0.19.3, JASP Team (2024)).

This analysis yielded a BF_{01} of 5.97, suggesting moderate evidence for the null hypothesis, indicating no difference in suspect-filler similarity between the two MTD methods ($Mdn = -.09$, 95% CI $[-.29, .11]$).

Figure 11

Bar Graph Depicting the Lineup Construction Method and Target Presence Interaction for Suspect-Filler Similarity Ratings



Overall Lineup Similarity

After looking at the filler-suspect similarity ratings at the level of individual face pairs, we investigated how individual filler-suspect similarity impacted the overall similarity ratings of

each lineup. First, we averaged the similarity ratings for each filler-suspect pairing. Each lineup constructed in Experiment 1b consisted of a suspect and five fillers. We then took the five relevant filler-suspect ratings for that lineup and averaged those to produce an overall lineup similarity score. Because lineups were the unit of analysis, and because each individual lineup was uniquely produced by a single lineup constructor, we were able to retain all fillers in this analysis.

Table 20

Mean Suspect-Filler Similarity Scores Across Construction Method

	BMTD		MTS		UMTD	
	TP	TA	TP	TA	TP	TA
Suspect-Filler Similarity	3.00 (0.48)	3.73 (0.57)	3.12 (0.59)	3.98 (0.60)	3.18 (0.47)	3.73 (0.65)

Note. Standard deviations for each mean are noted in parentheses.

In order to investigate whether lineup construction method and target presence influence overall lineup similarity, we submitted these lineup-wise similarity scores to an analysis of variance (ANOVA). The unit of analysis was the lineup; in total, there were 96 lineups with 16 in each cell of the design. We found a significant main effect of target presence on overall lineup similarity score, $F(1, 2) = 38.466, p < .001$, with a large effect size, $\eta^2 = 0.29$. Target absent lineups had higher similarity scores ($M = 3.815, SE = 0.082$) than target present lineups ($M = 3.099, SE = 0.082$). This means that fillers were more similar, on average, to innocent suspects than to guilty perpetrators. We did not find a significant main effect of lineup construction method on overall lineup member similarity, $F(2, 1) = 0.799, p = .45$, nor did we find an interaction effect, $F(2, 1) = 0.599, p = .55$. However, it is important to note that, because the

lineups were the unit of analysis, statistical power was lower than for the previous analysis. It is also worth noting that the pattern of means follows a similar pattern to those observed in the analysis of face pairs.

Discussion 3a

Overall, the findings of this study support our first hypothesis that the MTS method of lineup construction yields fillers who are rated as more similar to the suspect at both an individual pairing level and at the overall lineup level compared to the blinded version of the MTD method. This finding corroborates prior work in showing fillers in MTS lineups are consistently rated as more similar to the suspect than fillers in MTD lineups (Fitzgerald et al., 2015). It is also worth noting that despite MTS lineup constructors being instructed to select fillers “who look most similar to the suspect” (Chapter 2), mean suspect-filler similarity ratings for MTS lineups fell within the moderate range rather than approaching the highest levels of similarity.

However, our second hypothesis was not supported as the unblinded version of the MTD method performed more similarly to the blinded version than it did to the MTS method regarding filler-suspect similarity. This finding is surprising considering the moderate overlap in filler selection between lineups created using the unblinded MTD method and those created using the MTS method (see Chapter 2, Study 1b). It is also surprising as unblinded MTD lineup constructors reported using the image to make their filler selections (see Chapter 2, Study 1b). Therefore, top-down information processing would suggest lineup constructors in the unblinded version of the MTD method would be unable to disregard their knowledge of the suspect’s appearance, even while basing their filler selections on the perpetrator description, as existing knowledge impacts perception and decision making (Firestone & Scholl, 2014; Kinchla &

Wolfe, 1979). One possible explanation for this unexpected finding could be the difference in instructions provided to unblinded MTD lineup constructors and those using the MTS method. While lineup constructors in the unblinded MTD condition were instructed to select fillers “who best matched the description”, lineup constructors in the MTS condition were instructed to select fillers who “look most similar to the suspect”. Therefore, it appears unblinded MTD lineup constructors may have largely based their filler selection on the description rather than the suspect’s image, as instructed.

Another surprising finding of this study is the effect of target absence, as filler-suspect similarity was higher across all three construction methods for innocent suspects than for guilty perpetrators. Though it is unclear what is driving this effect of target presence. It is possible the guilty perpetrators used in this research may have had distinctive features which reduced their perceived similarity to the fillers presented alongside them, compared to innocent suspects. However, the data from the earlier mock witness task did not indicate any glaring differences between target present and target absent lineups which would suggest this. Thus, further research is needed to determine why suspect-filler similarity was higher for innocent suspects than for guilty perpetrators.

Study 3b

As previously discussed, filler selection method may indirectly impact suspect-filler similarity and lineup member fit to the description (Darling et al., 2008). However, while much research has been conducted on suspect-filler similarity, little has been done on lineup member fit to the description, such as how it may be influenced by filler selection methods and how it impacts lineup outcomes. Therefore, this research aims to be among the few to investigate how filler selection method impacts lineup member description-fit.

Method 3b

Participants. A sample of 100 participants were recruited through Prolific. Each participant viewed 97 culprit description-face pairings. Power analysis using G*Power v3.1 showed this was sufficient to detect a small effect size of $\eta^2 = 0.02$. Participants ranged in age from 18-74 years old ($M = 38.78$ and $SD = 13.30$) with the majority of participants being female (females = 62%, males = 36%, other/prefer not to respond = 2%). Participants were paid in accordance with Prolific guidelines, based on the study timing this resulted in £8.20/hour.

Design. This experiment followed a within-subject design. The independent variable of lineup construction had 3 levels: blinded MTD, unblinded MTD, and MTS; and the outcome variable being similarity ratings for the face and description pairings.

Materials.

Descriptions. The descriptions of the faces came from the high-quality descriptions produced in Study 1a.

Lineup member photographs. The photographs of the faces came from lineups created in Study 1b. In total, 179 faces were split into two lists and evaluated for description fit, including those of the guilty and innocent suspects. The faces were randomly assigned to one of the lists via a list randomiser. Participants were then randomly assigned one of the two lists, with the faces and corresponding descriptions shown in a random order through use of the randomisation function on Gorilla.

PI-20. Please refer to Study 3a (p. 123) for a description of the Prosopagnosia Index (PI-20). An example of the questionnaire can be found in *Appendix P*.

Procedure. This study was conducted online using the experiment-building platform Gorilla (www.gorilla.sc; Anwyl-Irvine, Massonnié, Flitton, Kirkham & Evershed, 2020) and participants

were recruited online from Prolific (www.prolific.co). Once participants clicked on the study link, they were redirected to an information screen, followed by an informed consent form. Once participants agreed to participate, they then viewed an instruction screen. The instruction screen advised participants they would be provided with a description and a picture of a face and asked to rate how well they believe the face matches the description provided on a 5-point scale, ranging from 1 (not at all) to 5 (matches very well). They were told to take their time while comparing the faces to the descriptions and that there was no right or wrong answer. They were to repeat this rating process for 97 pairings, with a short break midway through. Participants were shown the face and description pairings in a randomized order, using the randomize function on Gorilla. Upon rating the last of the face and description pairing, participants were asked to complete the PI-20. After completing the PI-20, participants were debriefed and told they could exit the survey. This study took approximately 15 minutes to complete.

Results 3b

We used the statistical software JASP (v0.19.3, JASP Team (2024)) to perform our statistical analyses. Out of the 100 participants in this study, 12 were excluded for scoring 65 or above on the PI-20 or for self-reporting a prosopagnosia diagnosis. First, we examined the description-fit ratings for each face and description pairing. For this analysis, face pairings formed the unit of analysis. Because we needed to assign each face pairing to a single construction method (MTS, blinded MTD or unblinded MTD), we excluded any fillers that were selected by lineup constructors across more than one construction method. For example, if a filler was chosen by both an MTS and a blinded MTD lineup constructor, that filler was excluded from this analysis. For this reason, 48 faces were excluded from this analysis.

As fillers in the blinded MTD condition were selected solely on the basis of their match to the perpetrator descriptions provided, we expected higher description-fit ratings than those fillers selected in the MTS condition. Additionally, we predicted the description-fit ratings for those faces selected in the unblinded MTD condition would more closely resemble those of the MTS method. This prediction was based on the self-reported strategies by lineup constructors in the unblinded MTD condition (Chapter 2) which reported using the suspect images to make their selections. To our surprise, we found that description-fit ratings were highest for unblinded MTD lineups, $M = 3.065$, $SD = 0.632$, compared to blinded MTD, $M = 2.876$, $SD = 0.503$, and MTS lineups, $M = 2.646$, $SD = 0.632$. This suggests that the lineup members selected for lineups created using the unblinded MTD method were a better match to the culprit's description than those selected using the blinded version of the MTD method or those selected using the MTS method.

We conducted a repeated measures analysis of variance (ANOVA) to investigate whether the variable of construction method influences the description-fit ratings for the face-description pairings, with lineup construction method as the main effect variable and description-fit rating as the dependent variable. We found a main effect of lineup construction method, $F(1.62, 140.57) = 36.83$, $p < .001$, with a large effect size, $\eta^2 = 0.30$. This is further supported by the post hoc test comparisons conducted using the Holm correction. This analysis showed that fillers selected using the blinded MTD method had higher description-fit ratings than those selected using the MTS method, $t(87) = 5.91$, $p < .001$. However, contrary to our hypothesis, fillers selected using the unblinded version of the MTD method had higher description-fit ratings than either the blinded MTD method, $t(87) = 4.12$, $p < .001$, or the MTS method, $t(87) = 7.03$, $p < .001$. In

addition to looking at the description-fit ratings at the stimulus level, we also investigated overall lineup description-fit.

Overall Lineup Description-Fit

After looking at the description-fit ratings at the level of individual face-description pairs, we wanted to investigate how individual description-fit impacted the overall description-fit ratings of each lineup. First, we averaged the description-fit ratings for each description-face pairing. We then took the six averaged description-fit ratings (six ratings because each face in the lineup, including the innocent and guilty suspect, was paired with the description associated with that stimulus set) and averaged those to get an overall lineup description-fit rating. As in Study 3a, because the lineups were the unit of analysis, and because each individual lineup was uniquely produced by a single lineup constructor, we were able to retain all fillers in this analysis. We then conducted an analysis of variance (ANOVA) to investigate whether lineup construction method and target presence influenced lineup similarity.

First, we examined the overall description-fit ratings for each lineup. We found description-fit ratings were higher for both blinded MTD, $M = 3.549$, $SD = 0.364$, and unblinded MTD lineups, $M = 3.538$, $SD = 0.367$, compared to MTS lineups, $M = 3.297$, $SD = 0.424$. This supported our hypothesis as we predicted the blinded MTD method would yield lineups with higher description-fit ratings compared to the MTS method, due to MTS lineup constructors selecting their fillers based on physical similarity to the suspect and blinded MTD lineup constructors selecting fillers based only on the perpetrator description provided. However, while it is not surprising that the blinded MTD method led to lineup members with better description-fit compared to lineup members in MTS lineups, it is surprising that the unblinded MTD method led to lineups more similar in description-fit ratings to the blinded MTD method than to the MTS

method. This finding suggests that lineup constructors were successfully using the MTD method whether blinded or not.

In order to investigate whether lineup construction method and target presence influence overall lineup description-fit, an analysis of variance (ANOVA) was conducted with the unit of analysis being the lineup, there were a total of 96 lineups with 32 lineups per each lineup construction method. We found a significant main effect of lineup construction method, $F(2,1) = 4.356, p = .016$, with a medium effect size of $\eta^2 = 0.08$. We conducted a post hoc test using the Holm correction to further investigate the main effect of construction method and found a significant difference in description-fit ratings between blinded MTD lineups and MTS lineups, $t(90) = 2.613, p = .03$, supporting our hypothesis that blinded MTD lineups would yield higher description-fit ratings compared to MTS lineups. Additionally, the post hoc test showed a significant difference in description-fit ratings between unblinded MTD lineups and MTS lineups, $t(90) = 2.496, p = .03$, with the unblinded MTD method yielding higher description-fit ratings compared to the MTS. There was no significant difference in description-fit ratings between blinded and unblinded MTD lineups, $t(90) = 0.117, p = .91$. To investigate this finding further, an exploratory Bayesian 2 (Lineup Construction Method: blinded MTD, unblinded MTD) x 2 (Target Presence: target present, target absent) ANOVA was conducted using JASP (v0.19.3, JASP Team (2024)). This analysis found support for the null hypothesis over the alternative, $BF_{01} = 3.89$, indicating no difference in description-fit between the two MTD methods. We also did not find a significant main effect of target presence on overall lineup member description-fit, $F(1, 2) = 2.168, p = .14$, nor an interaction effect, $F(2,1) = 0.522, p = .59$.

Discussion 3b

Overall, the findings of this study support our hypothesis that the blinded MTD method of lineup construction yields fillers who are rated as better matching the perpetrator description at both an individual stimulus level and at the overall lineup level compared to the MTS method. The findings of this study support previous research by Darling, Valentine, and Memon (2008) which found that fillers in MTD lineups were rated as matching the description better than fillers in MTS lineups. This finding is unsurprising as fillers in the MTD lineups were selected based on how well they matched the description, whereas fillers in MTS lineups were not.

However, our second hypothesis was not supported as the unblinded version of the MTD method performed more similarly to the blinded version of the method than it did to the MTS method in regard to description-fit, with it actually outperforming the blinded version of the MTD method at the individual pairing level. As in the case of 3a, this finding was unexpected as unblinded MTD lineup constructors reported using the image in some capacity during the filler selection process (see Chapter 2, Study 1b). This finding suggests that unblinded MTD lineup constructors may be giving more weight to details in the perpetrator description than to the information gained from the suspect's image.

General Discussion

The purpose of this study was to investigate how filler selection methods impact suspect-filler similarity and description-fit among lineup members, as these are both important facets of lineup fairness. We had four hypotheses: (1) the MTS method would yield lineups with higher suspect-filler similarity ratings compared to lineups created using the MTD method; (2) the unblinded version of the MTD method would yield suspect-filler similarity ratings more similar to those from MTS lineups than those from blinded MTD lineups; (3) the blinded version of the

MTD method would yield higher description-fit ratings than the MTS method; and (4) the unblinded version of the MTD method would yield description-fit ratings more similar to those seen in MTS lineups than those seen in blinded MTD lineups.

Only our hypotheses comparing the MTS method to the blinded MTD method were supported. That is, fillers selected using the MTS method were rated as more similar to the suspect than fillers selected using the blinded MTD method; conversely, fillers selected using the blinded MTD method were rated as better matches to the description than fillers selected using the MTS method. The findings of this study are in line with the findings of Darling, Valentine, and Memon (2008), who also found that fillers in MTS lineups resembled the suspect more than fillers in MTD lineups and, conversely, fillers in MTD lineups matched the culprit description better than fillers in MTS lineups. Tunnicliff & Clark (2000) also found that suspect-filler similarity was higher in MTS lineups than in MTD lineups, providing further support that MTS lineups consistently yield lineups higher in similarity than MTD lineups. Additionally, in a review of the literature Fitzgerald and colleagues (2015) found that filler selection method impacts suspect-filler similarity. They found that the MTS method consistently yields lineups higher in suspect-filler similarity than lineups created using the MTD method. This finding is further supported by their suggestion to interpret differences in similarity ratings between the two methods by substituting “moderate similarity” for the MTD method and “high similarity” with the MTS method (Fitzgerald et al., 2015).

These results are unsurprising, as the MTS method selects fillers based on how physically similar the fillers are to the suspect, whereas the MTD method selects fillers based on how well they match the features mentioned in the perpetrator description. Therefore, lineup members in MTS lineups should look more similar to one another, while lineup members in MTD lineups

may vary slightly from one another but still match in those features mentioned in the description. While these findings were expected, they provide useful information into how filler selection method may impact lineup fairness. As discussed in Chapter 3, the mock witness paradigm may only truly measure one aspect of lineup fairness, the degree to which lineup members match a description. It stands to reason, therefore, that lineup fairness metrics derived from a mock witness task would favour the MTD method over the MTS method (see also Fitzgerald, Oriet, & Price, 2015; Jones, 2021).

However, the findings of this study indicate that when considering another aspect of lineup fairness, perceived similarity, the MTS method may also lead to fair lineups. As suspect-filler similarity ratings measure how much the fillers resemble the suspect, they relate to lineup fairness estimates as these estimates are simply measures of how much the suspect stands out from the fillers in the lineup (Lee et al., 2025). Carlson, Gronlund, and Clark (2008) found that as suspect-filler similarity increased, lineup fairness also increased. As the results of our study found that lineup members in MTS lineups were rated as more similar to one another compared to lineup members in both versions of the MTD method, it may be possible that MTS lineups could be fairer than MTD lineups when measured by tasks other than the mock witness paradigm. This will be investigated further in the next chapter through the use of resultant fairness estimates.

In addition to investigating how lineup member similarity and description-fit differ in lineups created using the MTS method and the blinded MTD method of filler selection, we also compared the lineup member similarity and description-fit ratings of the blinded MTD method to the unblinded version of the method. Contrary to our predictions, the unblinded version of the MTD method performed more similarly to the blinded version of the method than to the MTS

method in terms of both filler-suspect similarity and lineup member description-fit. This finding was unexpected as unblinded MTD lineup constructors reported using the image alongside the perpetrator description to make their filler selection (see Chapter 2, Study 1b). Thus, the literature on top-down information processing would suggest lineup constructors in the unblinded version of the MTD method would have been hampered by having access to the suspect's photo, as existing knowledge impacts perception and decision making (Firestone & Scholl, 2014; Kinchla & Wolfe, 1979; Wells et al., 1998; Wells et al., 2020). Therefore, in theory the information gained from the suspect's image should have biased the fillers selected by unblinded MTD lineup constructors.

However, our findings do not support this. Instead, even though unblinded MTD lineup constructors had access to the suspect's image, perhaps they gave more weight to the features mentioned in the description of the perpetrator than they did those seen in the image of the suspect. This is interesting as research on blinding lineup administrators has consistently shown the negative impact an administrator's knowledge of the suspect can have on witness identifications (Wells et al., 2020). Therefore, as blinding lineup administrators is important in preventing administrators from biasing witness identification decisions, one would expect blinding MTD lineup constructors would lead to differences in lineup fairness measures when compared to an unblinded version of the method. However, the results of this study do not appear to support that prediction suggesting that the blinding of MTD lineup constructors may not be as important in filler selection and lineup construction as it is in lineup administration. With that said, these findings, taken along with those of the previous study on lineup fairness, suggest that when the MTD method is implemented properly, it may yield fairer lineups

compared to the MTS method while also decreasing the issue of high lineup member similarity and still ensuring good description-fit among lineup members.

Future Directions

While the results of this study appear to suggest the blinding of MTD lineup constructors may not be as important as previously believed, more research is needed to confirm how blinding MTD lineup constructors impacts various lineup outcome measures. To investigate this further, our final study will consist of a high-powered, traditional lineup study in which we will assess whether lineup construction method impacts witness accuracy, confidence, and discriminability. Additionally, we will also be looking at resultant lineup fairness and investigating how resultant fairness measures relate to mock witness fairness.

Chapter 5

Lineup Filler Selection Method and Lineup Outcomes

Decades of research have shown the importance of creating quality lineups with fillers who resemble the suspect to the point that the suspect does not stand out from the other lineup members. However, there has been much debate as how lineup constructors should select these fillers. As described in earlier chapters, there are two commonly used methods to select fillers for lineups: the MTS method and the MTD method. Lineup constructors using the MTS method select fillers based on their perceived similarity to the suspect. This method follows the guidelines set in the Police and Criminal Evidence Act 1984, Code D, (PACE), which states that all fillers should physically resemble the suspect (Malpass, Tredoux, & McQuiston-Surrett, 2007; Police and Criminal Evidence Act 1984, Code D, 2023). While the MTS method is the more widely used method (see Fitzgerald, Rubínová, & Juncu, 2021) and ensures lineup members resemble one another, the MTD method was proposed in response to issues with too high similarity among lineup members in MTS lineups (Wells et al., 1993). The MTD method addresses the issue of too high lineup member similarity by selecting fillers based on how well they match the features mentioned in the perpetrator description rather than on physical similarity to the suspect (Luus & Wells, 1991, Wells et al., 1993).

However, as argued in Chapter 1, the MTD method is rarely implemented in the manner it was originally proposed. The MTD method was first proposed by Wells and colleagues (1993) and involved matching fillers solely to the culprit description provided by eyewitnesses. Additionally, as a part of this proposed MTD method, Wells and colleagues (1993) selected new fillers and created a new lineup to match each witness description rather than combining all witness descriptions to create one modal description, as is often done in eyewitness research.

However, this method is now often used in tandem with the MTS method. For example, Fitzgerald and colleagues (2021) found that of the countries that recommend the use of the MTD method, many of them also recommend matching fillers to the suspect's appearance. An example of this could be that the description is used simply to narrow down the pool of fillers with the final filler selections being based on matching the suspect. Or the description could be used in the actual filler selection process alongside an image of the suspect, where fillers are selected based on how well they match the description *and* how closely they resemble the suspect.

Another way this method has been implemented is with the lineup constructor being unblinded to the suspect's appearance. In these instances, lineup constructors are following a strict MTD protocol (i.e. basing filler selections solely on the description) but also have seen the suspect in some capacity. Even using a strict MTD protocol, MTD lineup constructors could still be biased by their existing knowledge of the suspect. Therefore, these differences in the implementation of the MTD method could be undermining the effectiveness of this method. Thus, we proposed blinding MTD lineup constructors to the suspect's appearance, as we do lineup administrators to reduce information bias.

In Study 1, we created MTD and MTS lineups. The MTD lineups were created either by lineup constructors who were blinded to the appearance of the suspect (blinded MTD) or by lineup constructions who also had access to the suspect's image (unblinded MTD). In Chapters 3 and 4, we examined how filler selection method may impact lineup fairness, as assessed using a mock witness task, as well as suspect-filler similarity, and description fit. In this study, we will be comparing these lineups in a standard eyewitness experiment, which will enable us to examine how these filler selection methods impact lineup outcomes, discriminability, and resultant lineup fairness.

Filler Selection and Discriminability

As both the MTS method and the MTD method have their advantages and disadvantages, there has been much debate among researchers as to which method should be used by law enforcement to create lineups. On one hand, the MTS method ensures the suspect does not stand out among lineup members, as they were all selected based on their physical similarity to the suspect (Malpass, Tredoux, & McQuiston-Surrett, 2007). Additionally, the use of this method allows for lineup constructors to select quality fillers for a lineup even when the perpetrator description provided by the witness is too vague or is lacking sufficient detail to aid in filler selection (Lindsay et al., 1994). However, much research has shown that lineups created using the MTS method result in lineups higher in suspect-filler similarity than lineups created using the MTD method (Fitzgerald et al., 2015; Tunnicliff & Clark, 2000; Darling et al., 2008). This can become problematic because as similarity among lineup members increases, discriminability decreases, which can negatively impact the reliability and predictive utility of identification decisions (Carlson et al., 2019).

This decrease in discriminability is explained by two theories in the eyewitness literature: propitious heterogeneity (Luus & Wells, 1991; Wells et al., 1993) and the diagnostic feature detection (DFD) theory (Wixted & Mickes, 2014). The concept of propitious heterogeneity suggests that some distinctive details, or heterogeneity, across lineup members is needed in order for witnesses to be able to distinguish between guilty and innocent suspects (Wells et al., 1993). Luus and Wells (1991) argued that as long as fillers match all features provided in the perpetrator description, there is no reason to believe that increasing homogeneity in appearance among lineup members would provide better protection for an innocent suspect. Wells et al. (1993) provided an example of this, stating that lineup constructors seeking homogeneity in physical

features among lineup members could reach a point that the lineup consists of the suspect's clones, making the identification task nearly impossible for witnesses. Thus, they argued the importance of some "helpful differences", or propitious heterogeneity, among lineup members which would allow for witnesses to distinguish between guilty and innocent suspects (Luus & Wells, 1991; Wells et al., 1993).

The DFD theory is based on an important step in perceptual learning - the ability to discriminate between similar objects through the use of distinctive features (Wixted & Mickes, 2014). DFD theory applies this step of perceptual learning to eyewitness memory procedures, such as eyewitness identifications from suspect lineups, and states that discriminability is increased if eyewitnesses are able to identify and discount nondiagnostic features between lineup members, thereby allowing them to give more weight to any features that are diagnostic. For example, imagine an eyewitness provided the following perpetrator description to law enforcement: young, White male, brown curly hair, light-coloured eyes. Police then identify an individual who matches the general description (e.g. young, White male, dark hair) and present the suspect to the eyewitness using a photo lineup containing other young, White males, with dark hair. In this example, the descriptors regarding age, race, and hair colour would be nondiagnostic features as all lineup members share those features; while the descriptors of hair texture and eye colour may be diagnostic features as they were not included by lineup constructors creating the lineup, meaning these two features are less likely to be shared by guilty and innocent suspects and could aid witnesses in making their identification. The lineup task becomes increasingly difficult if the lineup members are very similar to one another as there are fewer diagnostic features for witnesses to use.

The MTS method is more likely than the MTD method to produce a lineup of highly

similar people, as there is no stopping rule beyond which the resemblance is too strong (Luus & Wells, 1991; Wells et al., 1993; Carlson et al., 2019). At the extreme end of the spectrum, hypothetical lineup of clones would make it nearly impossible for witnesses to make an identification as there would be very few, if any, diagnostic features for witnesses to identify and give weight to. Such a lineup is far more likely to be produced by an MTS filler selection strategy than by an MTD filler selection strategy. Therefore, the DFD theory offers a strong explanation regarding how filler selection method may impact discriminability.

As discussed in Chapter 1, a meta-analysis of the literature directly comparing the MTD method of filler selection to the MTS method found higher rates of correct IDs in lineups created using the MTD method than in lineups created using the MTS method. This finding may offer support for the MTD method yielding lineups with increased discriminability compared to the lineups created using the MTS method. The meta-analysis also showed higher rates of incorrect filler IDs for MTS lineups compared to the MTD lineups when the target was present. However, it found no significant differences in non-IDs between the two methods regardless of target presence, nor did it find a difference between the methods for filler IDs when the target was absent from the lineup. However, the meta-analysis did show higher rates of innocent suspect IDs for description-matched lineups in comparison to suspect-matched lineups. With that said, it is worth noting this finding was driven by a single, heavily weighted study, as the other studies included in the meta-analysis had small sample sizes. Therefore, additional, high-powered research is needed in order to determine how filler selection method impacts innocent suspect IDs. Additionally, many of the studies included in the meta-analysis pre-date the widespread usage of ROC analysis in the eyewitness identification literature. As a result of this, there are few studies which have directly compared the two filler selection methods using metrics such as area

under the curve (AUC) or partial area under the curve (pAUC).

ROC analysis is used to separate the contributions of discriminability from response bias (Gronlund & Neuschatz, 2014; Gronlund et al., 2014; Jones, 2021). ROC curves do this by plotting correct IDs against false IDs across various levels of confidence, or response bias, with the far-left point indicating the correct ID rate and false ID rate reported with the highest level of confidence and the far-right point indicating the correct ID rate and false ID rate made at all levels of confidence. The points in between represent the IDs made at varying cumulative levels of confidence, such as the point directly to the right of the far-left point will include the correct IDs and false IDs of the highest and second highest levels of confidence and so on until you get to the far-right point (Gronlund & Neuschatz, 2014). The ROC curve furthest away from the chance diagonal represents the condition with the best discriminability (Gronlund & Neuschatz, 2014; Gronlund et al., 2014). However, as ROC curves in lineup research are created using only suspect IDs (guilty perpetrators and innocent suspects); thus, excluding filler IDs and non-IDs, the resulting lineup ROC curves are truncated. This means a partial area under the curve (pAUC) must be calculated by cutting off the curve at the point that corresponds to the lowest observed false ID rate across the various research conditions to allow for comparison between the different ROC curves (Gronlund et al., 2014; Wixted & Mickes, 2018). In addition to assessing discriminability in studies comparing filler selection method (Carlson et al., 2019; Colloff et al., 2021), ROC analysis has been used to compare discriminability differences in fair versus biased lineups (Smith et al., 2016); the effect of pre-event instructions on participant performance in eyewitness experiments (Baldassari et al., 2023); and evaluating various eyewitness identification procedures (Gronlund et al., 2014; Wixted & Mickes, 2014). As the use of pAUC has been well established in the eyewitness identification literature, we will be using it in this

study to measure discriminability and will compare the pAUCs for each filler selection method to gain insight as to which method leads to increased discriminability.

One study included in the meta-analysis that used ROC analysis to investigate the impact of filler selection method on discriminability was conducted by Carlson and colleagues (2019). They found that the MTD method of filler selection yielded lineups with more propitious heterogeneity and higher empirical discriminability compared to lineups created using the MTS method of filler selection, as measured by pAUC. As such, one of the aims of this study is to replicate this finding.

Blinding Lineup Constructors and Discriminability

While existing research has found differences in discriminability between lineups created using the MTD method and lineups created using the MTS method, no research has been conducted investigating how blinding MTD lineup constructors to the suspect's appearance impacts eyewitness accuracy and confidence. However, as discussed in Chapter 1, the importance of blinding in lineup administration has been well documented, as an administrator's prior knowledge of the suspect may lead to them biasing a witness' identification decision. Research has consistently shown that lineup administrators who know the identity of the suspect are more likely to intentionally or unintentionally influence an eyewitness's identification through verbal or nonverbal cues (Wells et al., 1998; Wells et al., 2020; Kovera & Evelo, 2017). As research has established how blinding can reduce bias within the administration of lineups, it stands to reason blinding lineup constructors using the MTD could also reduce bias during the creation of the lineups.

If lineup constructors using the MTD method have prior knowledge of the suspect's identity, it may bias their decisions during the filler selection process. This could result in them

purposely selecting fillers who better match the suspect than they do the perpetrator description or influence their interpretation of the features mentioned in the description. Either scenario could result in a lineup with higher suspect-filler similarity than lineups created by only using the perpetrator description. Though these differences in suspect-filler similarity between lineups created using an unblinded and blinded version of the MTD method may be subtle, it could be enough to negatively impact discriminability. While there were no significant differences in lineup member similarity between blinded and unblinded MTD lineups (Chapter 4), the findings from Study 1b (Chapter 2) show that blinded and unblinded MTD lineup constructors selected different fillers. This could be due to unblinded MTD lineup constructors reporting their use of the suspect's image to select their fillers. As such, it is possible they selected fillers who resembled the suspect more than those selected by blinded MTD lineup constructors. Thus, this difference in fillers selected between the two MTD methods could lead to subtle differences in lineup member similarity which could in turn impact empirical discriminability.

Filler Selection and the Confidence-Accuracy Relationship

Another method of assessing witness accuracy is through assessing the confidence-accuracy relationship. The confidence-accuracy relationship is often depicted by a confidence-accuracy characteristic (CAC) analysis, which plots suspect identification accuracy at varying levels of confidence (Mickes, 2015). This analysis is useful in investigating variables that may affect eyewitness identification accuracy, such as exposure duration or delay between witnessed event and identification, by measuring the relationship between the subjective probability of a correct ID (witness confidence) and the objective probability of a correct ID (witness accuracy) (Mickes, 2015). Research has shown that the use of both ROC and CAC analyses in eyewitness research is important as they provide useful insight to different audiences, with ROC analysis

results proving useful to policymakers and CAC analysis results providing useful information to judges and jurors.

There is currently very little research into how filler selection method may impact the confidence-accuracy relationship (Mickes, 2015; Brewer & Wells, 2006; Wixted & Wells, 2017). Only one study included in the meta-analysis from Chapter 1 investigated this. In addition to using ROC analysis to investigate the impact of filler selection method on witness accuracy, Carlson and colleagues (2019) also conducted a confidence-accuracy characteristic (CAC) analysis. This analysis allowed them to investigate the reliability of the eyewitnesses' identifications across lineups created using different filler selection methods. They found a strong confidence-accuracy relationship regardless of which filler selection method was used; for both methods, witness confidence was indicative of witness accuracy and suspect IDs made at high confidence were accurate at a rate of 85% or higher. While much research has shown the importance of recording witness confidence data, little research has used CAC analysis to investigate how filler selection method impacts the confidence-accuracy relationship. Therefore, we will use CAC analysis in addition to ROC analysis to investigate how filler selection method impacts witness accuracy. This analysis will be exploratory as we have no strong grounds for making a priori predictions.

Resultant Lineup Fairness

While we are interested in investigating how filler selection method and blinding MTD lineup constructors impacts witness accuracy and confidence, we are also interested in the impact these variables have on lineup fairness. In Chapter 3, we tested lineup fairness through the use of the mock witness paradigm. However, Quigley-McBride and Wells (2021) suggested a better measure of lineup fairness – *resultant lineup size*. This method of testing lineup fairness uses the

data from actual eyewitness identification decisions from target absent trials to calculate traditional measures of lineup fairness, such as Tredoux's E, effective size, proportion of suspect selections (Quigley-McBride & Wells, 2021). Resultant lineup fairness uses only identification decisions from target absent lineups because in target present trials, selections of the suspect will be driven primarily by memory-based processes. In contrast, when the target is absent, selections of the innocent suspect are driven by a combination of guessing and lineup bias (Lee et al., 2022).

While the mock witness paradigm and resultant lineup size can both be used to measure lineup fairness, these measures do not always agree with each other. For example, lineups that appear to be relatively fair based on mock witness data may appear to be biased when actual eyewitness identification data is examined (Lee et al., 2025; Quigley-McBride & Wells, 2021). Lee and colleagues (2025) examined the concurrent validity (i.e. how well a measure correlates with other well-established measures) and predictive validity (i.e. how well a measure predicts performance) of several commonly used lineup-fairness measures. These measures include the mock witness paradigm and resultant lineup fairness and were investigated using mock crime and lineup identification data from existing research. They found that resultant lineup fairness estimates had weak or non-significant correlations with those from the mock witness paradigm. This is further supported by the finding that the mock-witness measures showed lineups in Evelo et al.'s (2018) study as being more biased than lineups in Mansour et al.'s studies (i.e. Mansour, 2020; Pennekamp et al., 2019; Horry et al., 2021), but the opposite was true for resultant fairness measures (Lee et al., 2025). Additionally, Lee and colleagues (2025) found that resultant lineup fairness measures were better predictors of eyewitness performance than mock-witness measures. The findings of this study suggest that the mock witness paradigm may not be the best

indicator of lineup fairness.

In the present study, we will use the observed eyewitness data from target-absent lineups to calculate resultant lineup fairness measures. While the mock witness results seen in Chapter 3 suggest that lineups created using the MTD method are fairer than lineups created using the MTS method, no studies have previously looked at the impact of filler selection method on resultant lineup fairness estimates. Therefore, this will be an exploratory analysis to investigate how filler selection method impacts resultant lineup fairness.

Study aims

The purpose of this study is to investigate how filler selection methods and the blinding of lineup constructors impact witness discriminability, confidence, and resultant lineup fairness. This will be done using a traditional lineup study in which participants will view a mock crime video and then be shown a target present or target absent lineup created using one of the three lineup construction methods (MTS, blinded MTD, or unblinded MTD). Based on the existing literature, we hypothesised that discriminability would be higher for blinded MTD lineups than for MTS lineups. We also hypothesised that discriminability would be higher for blinded MTD lineups than for unblinded MTD lineups. These hypotheses were pre-registered on the Open Science Framework (<https://doi.org/10.17605/OSF.IO/JB8FZ>).

In addition to the two pre-registered hypotheses above, we tested several other, non-pre-registered hypotheses. Based on the findings from the meta-analysis conducted earlier in my studies (Chapter 1) and the lineup fairness chapter (Chapter 2), we predicted that MTD lineups, whether blinded or unblinded, would yield higher correct ID rates compared to the MTS lineups. We also predicted MTS lineups would yield higher innocent suspect ID rates compared to MTD lineups. Additionally, we predicted no differences in filler IDs or non-IDs/correct rejections

between MTD and MTS lineups. While we did not have any directional hypotheses regarding the CAC analysis and resultant fairness, we wanted to explore the impact filler selection method may have on these measures. We were also interested in investigating how blinding MTD lineup constructors would impact the confidence-accuracy relationship and resultant lineup fairness.

Method

Participants. In our pre-registered stopping rule, we stated we would recruit 1200 participants or terminate recruitment on 30 June 2025. In order to conduct the planned analyses for this study, we calculated we needed a minimum of 200 observations per lineup. This was based on existing eyewitness identification research utilising ROC analysis (e.g. Carlson et al., 2019). However, by the recruitment deadline (30 June 2025), we had recruited a sample of 843 participants through online methods, with 66% of participants recruited through Prolific, 33% of participants recruited through the university participant pool, and the remaining 1% recruited through social media. Participants ranged in age from 18-77 years old ($M = 30.39$ and $SD = 12.17$) with the majority of participants being female (females = 58.96%, males = 39.26%, other/prefer not to respond = 1.78%). University students were compensated with course credits and those recruited from Prolific were paid in accordance with Prolific guidelines. Based on the study duration, this resulted in a rate of £8.70/hour.

Design. This experiment followed a 3 (Lineup construction method: MTS, unblinded MTD, blinded MTD) x 2 (Target presence: target present versus target absent) design. Each participant viewed four lineups, one for each stimulus video. Each lineup was randomly sampled from the six cells of the design.

Materials

Videos. The materials used in this study were created in previous experiments. The four staged

mock crime videos used in this study came from research conducted by Molinaro et al (2021). Based off the identification data from Molinaro et al (2021), we decided to use the close-view, short videos. These videos were lower in quality compared to the high-quality videos used in Study 1a, but better in quality than the low-quality videos previously used. These videos were approximately 15 seconds in length, with a close-up view of the perpetrator. These stimuli were used with permission and are openly available via a Creative Commons license (CC-BY 4.0), stimulus materials can be located through Open Science Framework (https://osf.io/dz5pf/?view_only=2b41a6ce5c4d43e0abb08ea1f894d6b2).

Lineups. The lineups used are the same 24 randomly selected lineups used in the mock witness task in Chapter 3. Though 96 six-person lineups were originally created in Study 1b, we determined it would not be possible to recruit enough participants to view all 96 lineups and achieve sufficient power. These photographic lineups were presented simultaneously, displayed in two rows of three.

PI-20. Please refer to Study 3a (p. 123) for a description of the Prosopagnosia Index (PI-20). An example of the questionnaire can be found in *Appendix P*.

Procedure. This study was conducted online using the experiment-building platform Gorilla (www.gorilla.sc; Anwyl-Irvine, Massonnié, Flitton, Kirkham & Evershed, 2020). Once participants clicked on the study link, they were redirected to an information screen, followed by an informed consent form. Once participants agreed to participate, they then viewed an instruction screen. The instruction screen advised participants the following: “In a moment, you will see a short video, approximately 15 seconds in length. Please watch the video in its entirety before proceeding to the next page. After watching the video, you will be asked to find as many words as possible in 60 seconds from a word search. You will then be shown a lineup which *may*

or may not contain the individual seen in the video. Please take your time in making your identification. After your identification decision, you will also be asked to rate how confident you are in your decision.”

Participants were informed beforehand they would view a lineup in efforts to remain consistent across all four trials, as once the first trial was completed, participants would understand the nature of the study and expect the following lineups. Additionally, research on pre-event instructions has shown participants who were provided eyewitness pre-event instructions did not differ in their lineup identifications or confidence-accuracy ratings compared to participants who were provided non-specific pre-event instructions (Baldassari et al., 2023). As this finding shows no real effect on identification decisions or witness confidence, we decided there would be little harm in providing eyewitness pre-event instructions to the participants in this study.

Upon completing the video and word search, participants were randomly assigned one of the six possible lineups that were created for that stimulus video and asked to make their identification and confidence rating (see Figure 12). Participants indicated their decision by selecting the multiple-choice answer that corresponded to the individual’s position in the lineup or selecting the response “None of these / individual is not present” (see Appendix Q for example).

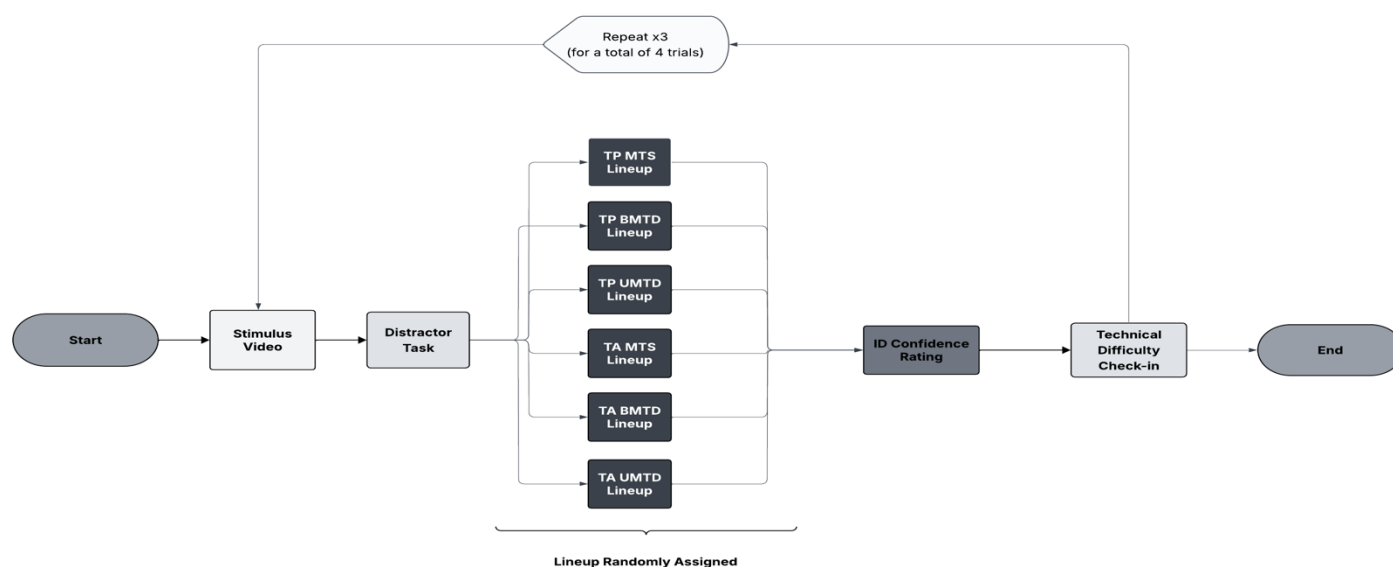
Confidence ratings were made on a sliding scale from 0 (not confident) to 100 (extremely confident) (see Appendix R for example). Upon making their identification and confidence rating, participants were asked two ‘yes or no’ questions to indicate whether they had experienced any technical difficulties when viewing the video or lineup photographs.

This process was then repeated for each of the remaining three stimulus sets, which were presented to participants in a randomised order. Upon making the last identification and confidence rating, participants were asked a series of questions to ensure that there were no

technical difficulties and that participants did not use any aids (i.e. screenshots of perpetrators in video) when making their identifications. They were also asked to answer the PI-20 questionnaire.² After completing the PI-20, participants were debriefed and told they could exit out of the survey. The study took approximately 15 minutes to complete.

Figure 12

Trial Process Flowchart



Results

This study and analysis plan was preregistered on the Open Science Framework (<https://doi.org/10.17605/OSF.IO/JB8FZ>). To investigate the impact of lineup construction method on witness accuracy and lineup fairness, we conducted several analyses including: a multi-level logistic regression to explore the impact of lineup construction method, target-presence, and their interaction on decision outcomes; CAC analysis to explore the impact of

² PI-20 was not used to screen out participants for this study.

lineup construction method on the confidence-accuracy relationship; ROC analysis to calculate partial Area Under the Curve, as well as pairwise comparisons to compare discriminability between the blinded and unblinded MTD conditions and between the blinded MTD and MTS conditions; and calculated resultant lineup fairness measures to explore the impact of lineup construction method on lineup fairness.

As stated in the pre-registration, individual trials were excluded if participants responded ‘no’ to one or both of the following questions: “were you able to watch the entire video without issue?” and “were you able to see the image containing the lineup of individuals for the video?” This resulted in the removal of 168 trials from the analysis.

Regression Analysis

Suspect Identifications

We used the statistical software JASP (v0.18.3, JASP Team (2024)) to perform our statistical analyses for each identification type. We conducted three separate mixed effects logistic regressions in order to investigate whether lineup construction method and target presence impact the different identification decisions (suspect IDs, filler IDs, non-IDs/correct rejections) made by eyewitnesses. Each of the models had the following structure: the identification decisions (suspect IDs, filler IDs, and non-IDs/correct rejections) served as the outcome variables, and lineup construction method, target presence, and their interaction as fixed effects. Stimulus set was included as a random intercept. These analyses used sum, or effect, contrast coding to allow us to investigate any main or interaction effects (Brehm & Alday, 2022).

First, we examined the proportion of suspect IDs made from each lineup construction method (*Table 21*). We found a main effect of target presence on suspect identifications, $\chi^2 = (1, N = 3358) = 854.22, p < .001$. Unsurprisingly, contrast analysis comparing the estimated

marginal means of target present and target absent lineups revealed guilty perpetrators ($M = 0.55$, $SE = 0.06$) were more likely to be selected by eyewitnesses than innocent suspects ($M = 0.09$, $SE = 0.02$), $\beta = 0.46$, $SE = 0.04$, $z = 11.17$, $p < .001$. We did not find a significant main effect of construction method $\chi^2 = (2, N = 3358) = 1.06$, $p = .59$. However, the interaction between lineup construction method and target presence was statistically significant, $\chi^2 = (2, N = 3358) = 19.26$, $p < .001$. Further analysis of the interaction using contrast analysis showed that the effect of lineup construction was different for target present lineups compared to target absent lineups. For target present lineups, the blinded MTD lineups yielded more correct suspect IDs ($M = 0.61$, $SE = 0.05$) compared to both MTS ($M = 0.51$, $SE = 0.05$), $\beta = 0.10$, $SE = 0.03$, $z = 3.25$, $p = .001$, and unblinded MTD lineups ($M = 0.52$, $SE = 0.05$), $\beta = 0.09$, $SE = 0.03$, $z = 3.08$, $p = .002$. There was no difference in correct suspect IDs between MTS and unblinded MTD lineups, $\beta = 0.003$, $SE = 0.03$, $z = 0.09$, $p = .93$. However, this pattern changes for target absent lineups, with innocent suspect IDs lower for blinded MTD lineups ($M = 0.06$, $SE = 0.02$) compared to MTS lineups ($M = 0.10$, $SE = 0.03$), $\beta = -0.04$, $SE = 0.02$, $z = -2.31$, $p = .02$. There were no significant differences in innocent suspect IDs between blinded MTD and unblinded MTD lineups ($M = 0.09$, $SE = 0.02$), $\beta = -0.03$, $SE = 0.02$, $z = -1.62$, $p = .10$, or between unblinded MTD lineups and MTS lineups, $\beta = -0.02$, $SE = 0.02$, $z = -1.05$, $p = .29$.

To confirm these observations, we created separate binary logistic regressions for target present and target absent lineups. The model structure was as follows: suspect IDs as the outcome variable and lineup construction method as the predictor variable. The model was significant for target present lineups, $\chi^2 = (1657) = 12.34$, $p = .002$, with correct IDs being more likely to occur than in blinded MTD lineups ($OR = 1.53$, 95%CI [1.29, 1.81]) than MTS ($OR = 0.69$, 95%CI [0.55, 0.88]) or unblinded MTD lineups ($OR = 0.69$, 95%CI [0.54, 0.88]). The

model was also significant for target absent lineups, $\chi^2 = (1657) = 8.24, p = .02$, with innocent suspect IDs more likely to occur in MTS lineups (OR = 1.78, 95%CI [1.19, 2.66]) than either unblinded MTD (OR = 1.42, 95%CI [0.93, 2.18]) or blinded MTD lineups (OR = 0.08, 95%CI [0.06, 0.11]).

Table 21

Identification Rates per Lineup Construction Method and Target Presence

Filler Selection Method	Target	Suspect ID rate	Filler ID rate	Non-ID rate
MTS	TP	.515 (299/581)	.272 (158/581)	.215 (125/581)
	TA	.122 (73/596)	.413 (246/596)	.466 (278/596)
BMTD	TP	.604 (336/556)	.196 (109/556)	.203 (113/556)
	TA	.073 (41/564)	.445 (251/564)	.480 (271/564)
UMTD	TP	.512 (268/523)	.283 (148/523)	.205 (107/523)
	TA	.100 (54/538)	.470 (253/538)	.431 (232/538)

Filler Identifications

We then examined the proportion of filler IDs made from each lineup construction method (*Table 21*). We found a main effect of lineup construction method on filler identifications, $\chi^2 = (2, N = 3358) = 10.06, p = .007$. Further analysis of the main effect using contrast analysis showed fillers were significantly more likely to be selected from unblinded MTD lineups ($M = 0.37, SE = 0.04$) than from blinded MTD lineups ($M = 0.31, SE = 0.04$), $\beta = 0.06, SE = 0.02, z = 2.82, p = .005$. There were no significant differences in filler identifications between unblinded MTD lineups and MTS lineups ($M = 0.33, SE = 0.04$), $\beta = 0.03, SE = 0.02, z = 1.65, p = .10$, or between blinded MTD lineups and MTS lineups, $\beta = -0.02, SE = 0.02, z = -1.24, p = .22$. We also found a main effect of target presence on filler identifications, $\chi^2 = (1, N = 3358) = 143.70, p < .001$. Contrast analysis showed fillers were more likely to be selected

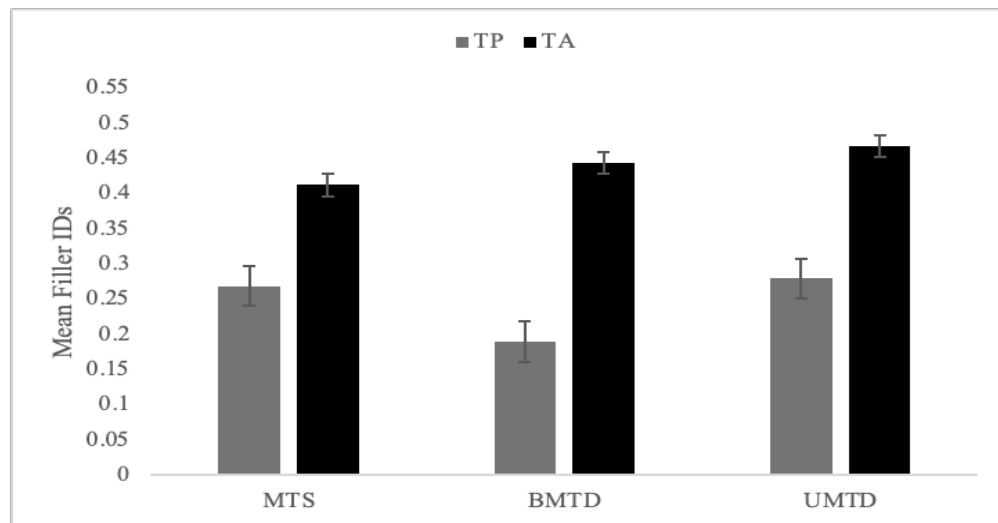
when the target was absent from the lineup, $M = 0.42$, $SE = 0.04$, than when the target was present in the lineup, $M = 0.24$, $SE = 0.03$, $\beta = 0.20$, $SE = 0.02$, $z = 10.21$, $p < .001$.

Additionally, we found a significant interaction between lineup construction method and target presence on the likelihood of a filler being selected from a lineup, $\chi^2 = (2, N = 3358) = 10.00$, $p = .007$ (*Figure 13*). Further analysis of the interaction using contrast analysis showed fillers were less likely to be selected from blinded MTD lineups ($M = 0.19$, $SE = 0.03$) compared to unblinded MTD lineups ($M = 0.28$, $SE = 0.04$), $\beta = -0.09$, $SE = 0.03$, $z = -3.33$, $p < .001$, and MTS lineups ($M = 0.27$, $SE = 0.03$), $\beta = -0.08$, $SE = 0.03$, $z = -3.02$, $p = .003$, when the target was present. There was no difference between unblinded MTD and MTS lineups in filler IDs when the target was present, $\beta = 0.01$, $SE = 0.03$, $z = 0.43$, $p = .67$. However, there was no significant difference in filler IDs between the three lineup construction methods when the target was absent from the lineup. The following contrasts are reported for completeness: blinded MTD lineups ($M = 0.44$, $SE = 0.05$) and MTS lineups ($M = 0.41$, $SE = 0.05$), $\beta = -0.03$, $SE = 0.03$, $z = -0.99$, $p = .32$; blinded MTD lineups and unblinded MTD lineups ($M = 0.47$, $SE = 0.05$), $\beta = 0.03$, $SE = 0.03$, $z = 0.86$, $p = .39$; and MTS and unblinded MTD lineups, $\beta = 0.05$, $SE = 0.03$, $z = 1.85$, $p = .06$.

This was supported by separate binary logistic regressions conducted on target present and target absent lineups. The model structure was as follows: filler IDs as the outcome variable and lineup construction method as the predictor variable. The model was significant for target present lineups, $\chi^2 = (1657) = 13.52$, $p = .001$, with filler IDs being more likely to occur in unblinded MTD lineups ($OR = 1.62$, 95%CI [1.22, 2.15]) than MTS ($OR = 1.53$, 95%CI [1.16, 2.02]) or blinded MTD lineups ($OR = 0.24$, 95%CI [0.20, 0.30]). The model was not significant for target absent lineups, $\chi^2 = (1657) = 3.83$, $p = .15$.

Figure 13

Interaction Between Lineup Construction Method and Target Presence for Filler IDs



Non-Identifications or Lineup Rejections

We also examined the proportion of non-IDs, or rejected lineups, for each lineup construction method (*Table 21*). We found a main effect of target presence on non-identifications, $\chi^2 = (1, N = 3358) = 246.69, p < .001$. Contrast analysis showed that when the target was absent from the lineup, $M = .46, SE = .03$, witnesses were more likely to make a non-identification decision than when the target was present, $M = .20, SE = .02, \beta = 0.25, SE = 0.02, z = 13.62, p < .001$. We did not find a significant main effect of construction method $\chi^2 = (2, N = 3358) = 1.57, p = .45$, nor a significant interaction $\chi^2 = (2, N = 3358) = 1.25, p = .54$.

ROC Analysis

To investigate whether filler selection method impacts empirical discriminability, or the witness's ability to discriminate between guilty perpetrators and innocent suspects, we conducted an ROC analysis. This analysis separates the contributions of discriminability from response bias by creating ROC curves for each condition (Gronlund & Neuschatz, 2014; Gronlund et al., 2014; Jones, 2021). ROC curves plot cumulative HR and FAR across various levels of confidence, or

response bias, with the far-left point indicating the HR and FAR reported with the highest level of confidence and the far-right point indicating the HR and FAR made at any level of confidence. The points in between represent the IDs made at varying cumulative levels of confidence, such as the point directly to the right of the far-left point will include the HR and FAR of the highest and second highest levels of confidence and so on until you get to the far-right point. The ROC curve furthest away from the chance diagonal represents the condition with the best discriminability (Gronlund & Neuschatz, 2014; Gronlund et al., 2014).

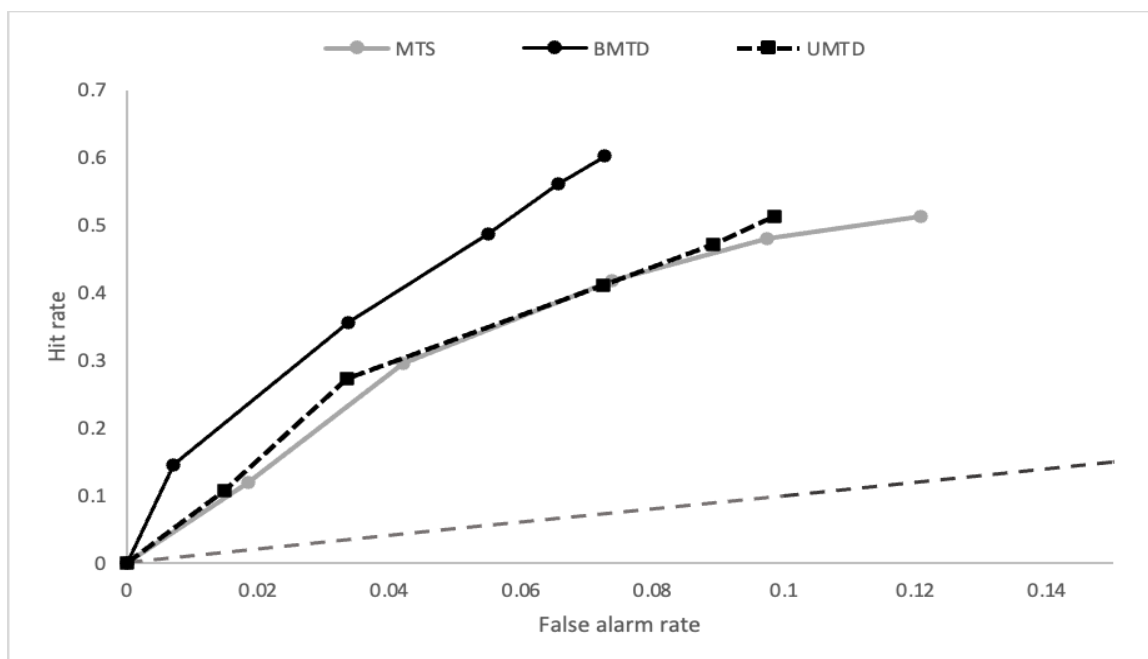
However, ROC curves in lineup research are created using only suspect IDs (guilty perpetrators and innocent suspects), with filler IDs and non-IDs being excluded. This exclusion of filler IDs results in truncated lineup ROC curves, which means a partial area under the curve (pAUC) must be calculated by cutting off the curve at the point that corresponds to the lowest observed FAR across the three conditions, for the purpose of this study that lowest observed FAR was 0.0727 (Gronlund et al., 2014; Wixted & Mickes, 2018). We calculated partial Area Under the Curve using the pROC package for R (Robin et al., 2011) to determine whether the blinded MTD lineups would yield better empirical discriminability compared to both MTS and unblinded MTD lineups. We compared the blinded MTD ROC curve with the MTS curve, collapsing over stimulus (specificity: $1 - 0.0727 = 0.927$, see *Figure 14*). As predicted, the blinded MTD lineups ($pAUC = .026$, $[.020-.032]$) yielded better discriminability than the MTS lineups ($pAUC = .017$, $[.013-.022]$), $D = -2.198$, $p = .03$.

We also compared the blinded MTD curve with the unblinded MTD curve. We found that blinding the lineup constructor in MTD lineups ($pAUC = .026$, $[.020-.032]$) increased discriminability compared to the unblinded MTD lineups ($pAUC = .018$, $CI [.013-.023]$), $D = 2.018$, $p = .04$, which supported our hypothesis. For completeness, we also compared the ROC

curves between the MTS lineups and the unblinded MTD lineups and found no significant difference between them, $D = -0.254$, $p = .79$. The ROC curves for each of the three lineup construction methods are depicted in the figure below.

Figure 14

ROC Curves for Each Lineup Construction Method Collapsed Across Stimuli



Confidence-Accuracy Characteristic

In addition to running multi-level logistic regressions for each identification type and ROC analysis, we wanted to investigate whether lineup construction method had an impact on the confidence-accuracy relationship. Confidence-accuracy characteristic (CAC) analysis is a form of calibration analysis and measures the relationship between confidence and accuracy by plotting suspect identification accuracy at each level of confidence for each lineup condition (Mickes, 2015). For example, if there are 100 suspect IDs made with 90-100% confidence, and 95 of those IDs are correct, then suspect ID accuracy at this confidence level would be .95. For this study, confidence was separated into five levels, or bins: 0-29%, 30-49%, 50-69%, 70-89%,

90-100%. At each level of confidence, we used a chi-square test to compare the confidence-accuracy relationship across all three lineup construction methods.

We found a significant relationship between construction method and accuracy only at the lowest confidence level (0-29%), $\chi^2 (2, N = 86) = 6.85, p = .03$. For completeness, the results of the other chi-squared tests are as follows: the second lowest confidence level (30-49%), $\chi^2 (2, N = 138) = 3.42, p = .18$; the next confidence level (50-69%), $\chi^2 (2, N = 268) = 2.29, p = .32$; the second highest confidence level (70-89%), $\chi^2 (2, N = 346) = 0.15, p = .93$; and the highest confidence level (90-100%), $\chi^2 (2, N = 229) = 4.32, p = .11$.

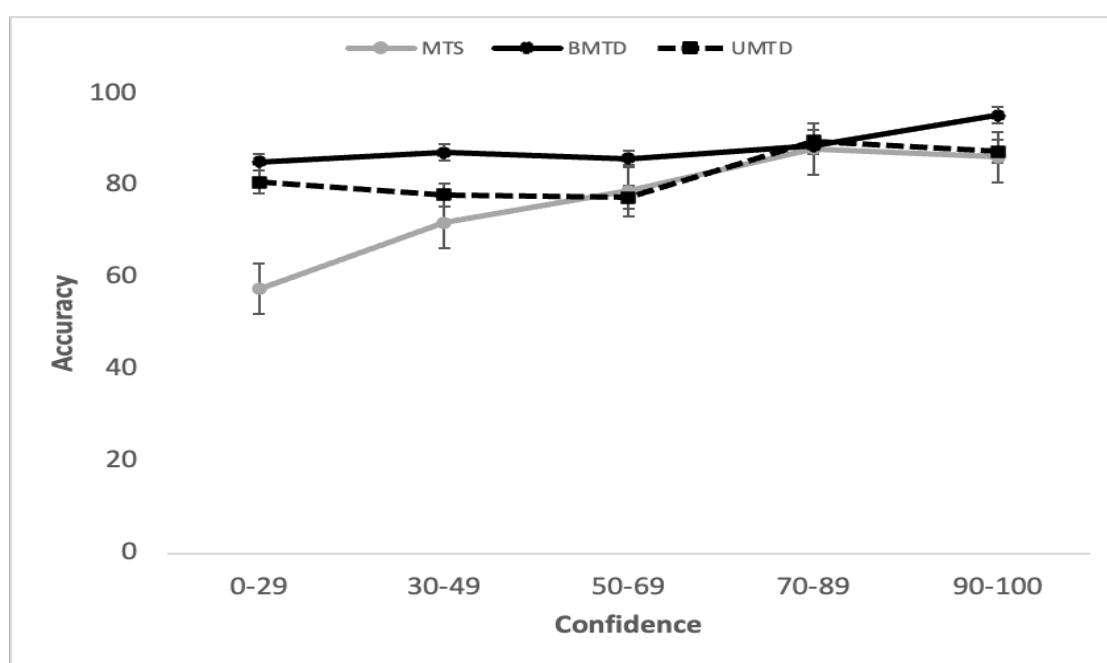
Upon further investigation, this impact on the confidence-accuracy relationship seems to be driven by the significant difference in confidence-accuracy between the blinded MTD and MTS lineups at the lowest confidence level (0-29%), $\chi^2 (1, N = 60) = 5.39, p = .02$, and at the highest confidence level (90-100%), $\chi^2 (1, N = 165) = 4.08, p = .04$. All other comparisons between blinded MTD and MTS lineups were not significant, as were all comparisons between blinded MTD and unblinded MTD lineups and between unblinded MTD lineups and MTS lineups. Exploratory chi-square analyses looking at identifications made only at the 100% confidence point across the three construction methods were also conducted but were not significant.

While the above analyses suggest construction method may impact witness accuracy at the lowest level of confidence, a CAC analysis was conducted to further investigate the confidence-accuracy relationship. The findings of the CAC analysis suggest the presence of a confidence-accuracy relationship for MTS lineups, such that as witness identification confidence increases so does identification accuracy. However, the confidence-accuracy relationship for blinded MTD and unblinded MTD lineups does not appear to be as straightforward. This is due

to the finding that identification accuracy for both MTD lineup construction methods was quite high, around the 80% mark, even at the lowest confidence level. Therefore, while identification accuracy did increase as witness confidence increased, the changes were subtle compared to those seen in MTS lineups. The figure below depicts the CAC for each of the three lineup construction methods (*Figure 15*).

Figure 15

CAC for Each Lineup Construction Method Collapsed Across Stimuli



Note: Error bars denote SEs.

Resultant Lineup Fairness

The statistical analyses performed to investigate lineup fairness were performed using R Statistical Software (v4.1.2; R Core Team 2021) via the statistical package *r4lineups* (v0.1.1; Tredoux and Naylor, 2018). This allowed us to calculate the outcome variables used to assess lineup fairness such as Tredoux's E, effective size, and the proportion at which the suspect was chosen from the lineup by eyewitnesses (Tredoux, 1998). The lineup fairness estimates in this

chapter were calculated using the observed lineup data, rather than mock witness data as used in Chapter 3. Lineups are considered to be fair when they have an effect size or Tredoux's E value closer to the nominal size of the lineup, as this suggests each lineup member was chosen at a similar rate. The lineup fairness measures of Tredoux's E and effective size were calculated for all 12 of the target absent lineups using the identifications made by choosers, as recommended by Quigley-McBride and Wells (2021). Target present lineups are not used to calculate resultant lineup fairness, as responses to target-present lineups would not only be influenced by the fairness of the lineup but also by the participants' memory for the target (Lee et al., 2025).

Overall, the lineups constructed using the blinded MTD method yielded Tredoux's E and effective size values closer to the nominal size of the lineup than lineups constructed using either the MTS or unblinded MTD lineups. This suggests that the lineups constructed using the blinded MTD method were fairer than lineups constructed using either the MTS or the unblinded version of the MTD method. This is further supported as the effective sizes for the lineups created by the three construction methods followed a similar pattern to that seen with Tredoux's E. While lineups constructed using both MTD methods yielded similar effective sizes, the lineup constructed using the blinded MTD effective size was slightly higher than that of the lineups constructed using the unblinded MTD method, with the lineups constructed using the MTS method still having the lowest effective sizes. The table displaying the descriptive statistics for Tredoux's E and effective size for each construction method can be seen below (*Table 22*).

Table 22

Mean Resultant Tredoux's E and Effective Size Values for Each Lineup Construction Method

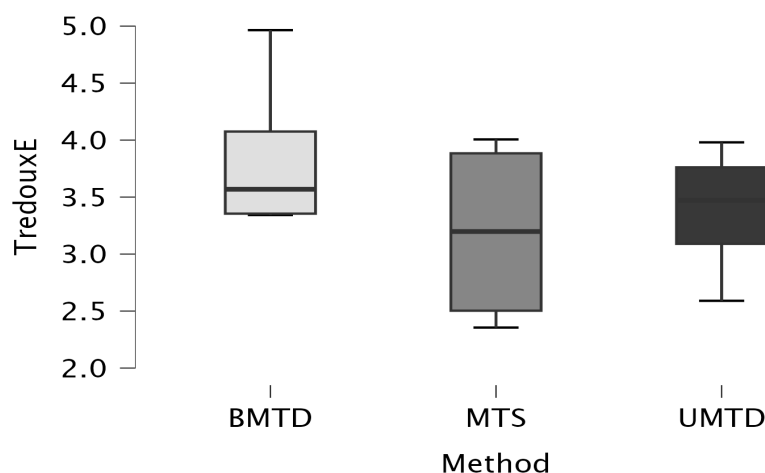
	Tredoux's E			Effective Size		
	BMTD	MTS	UMTD	BMTD	MTS	UMTD
Mean	3.861	3.189	3.379	4.344	3.941	4.218
95% CI Mean						
Upper	5.075	4.550	4.340	5.364	4.768	5.016
95% CI Mean						
Lower	2.647	1.828	2.418	3.323	3.115	3.419
Std. Deviation	0.763	0.855	0.604	0.641	0.519	0.502

The results from this study follow a similar pattern to those seen in the mock witness data (Chapter 3). Such that for both studies the MTS method yielded lineups with the lowest Tredoux's E (Figure 16) and effective size (Figure 17) values. These findings suggest that the blinded MTD method yields fairer lineups compared to the MTS method regardless of whether mock witnesses or actual eyewitnesses are making the identifications. This suggests that lineups created using the MTS method of lineup construction are more likely to be biased against the suspect, which could in turn impact witness accuracy and lead to wrongful convictions in real world situations.

However, there was a deviation in the pattern between the mock witness and resultant fairness data when looking at how blinding may impact lineup fairness. In Chapter 3, we found that when the target was absent, the unblinded version of the MTD method yielded lineups with slightly higher Tredoux's E values than lineups created using either the blinded MTD method or the MTS method. With that said, the results of this present study suggest that lineups created using the blinded MTD method led to higher Tredoux's E values than those created using the unblinded version of the method.

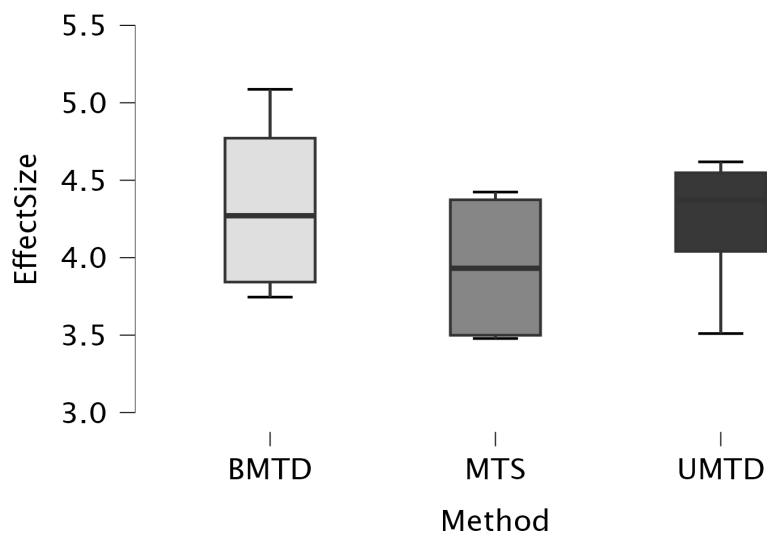
Figure 16

Boxplot Depicting the Mean Resultant Tredoux's E Values for Each Lineup Construction Method

**Figure 17**

Boxplot Depicting the Mean Resultant Effective Size Values for Each Lineup Construction

Method



In addition to examining Tredoux's E and effective size, we also looked at lineup fairness measures such as proportion of suspect selections, resultant suspect bias, and resultant defendant bias. These measures give insight into how often the innocent suspect was selected by witnesses, as well as if the identifications were skewed toward or away from the innocent suspect. The

innocent suspect was more likely to be selected from lineups created using the MTS method than from lineups created using either version of the MTD method (*Table 23*). Additionally, lineups created using the MTS method were higher in suspect bias than lineups created using either MTD method, with the blinded version of the MTD method having the least amount of suspect bias, further suggesting the identifications were skewed towards the suspect in lineups constructed using the MTS method.

Table 23

*Mean Proportion Suspect Selections, Resultant Suspect Bias, and Resultant Defendant Bias
Values for Each Lineup Construction Method*

	Proportion Suspect Selections			Suspect Bias			Defendant Bias		
	BMTD	MTS	UMTD	BMTD	MTS	UMTD	BMTD	MTS	UMTD
Mean	0.140	0.233	0.188	-0.639	1.526	0.303	-1.839	-0.419	-1.213
95% CI Upper	0.389	0.453	0.452	4.658	6.557	5.957	3.430	3.189	4.392
95% CI Lower	-0.110	0.014	-0.076	-5.936	-3.504	-5.351	-7.109	-4.027	-6.818
Std. Dev	0.157	0.138	0.166	3.329	3.161	3.553	3.312	2.268	3.522

Discussion

The purpose of this study was to investigate the impact of lineup construction method on witness discriminability, confidence, and resultant lineup fairness. We hypothesised that lineups created using the blinded MTD method would yield better empirical discriminability compared to lineups created using the MTS method. This hypothesis was supported by an ROC analysis. This finding offers support for two well-known theories from the eyewitness identification literature, propitious heterogeneity and the diagnostic feature-detection theory, which predict that as similarity among lineup members increases, discriminability decreases (Wells et al., 1993; Wixted & Mickes, 2014).

As demonstrated in the previous study (Study 3a), the blinded MTD lineups used in this study yielded lower suspect-filler similarity ratings compared to the MTS lineups. This finding, in combination with the results of the ROC analysis, suggest that as suspect-filler similarity increases, discriminability decreases. Our findings replicate those of Carlson and colleagues (2019), who found that lineups created with description-matched fillers were associated with higher empirical discriminability than lineups created with suspect-matched fillers.

We also hypothesised that lineups created using the blinded MTD method would yield better empirical discriminability than lineups created using the unblinded MTD method. This was supported by the ROC analysis. While no significant differences were seen between the two versions of the MTD method in my earlier studies, it appears that blinding lineup constructors using the MTD method impacts witness discriminability. Blinding MTD lineup constructors may have prevented their filler selections from being biased by existing knowledge of the suspect's appearance, increasing the effectiveness of the method and adding important variation among lineup members. As the literature on top-down processing suggests prior knowledge and experiences can influence perception and decision making, it stands to reason knowing what the suspect looks like would impact the decisions made by lineup constructors during the filler selection process (Firestone & Scholl, 2014; Kinchla & Wolfe, 1979). With that said, all but one of the unblinded MTD lineup constructors reported using the image of the suspect to make their filler selections (Chapter 2, Study 1b). Therefore, this may have led to unblinded MTD lineup constructors unconsciously selecting fillers who more closely matched both the suspect and description, resulting in diminished discriminability.

This brings to question what is driving these differences in discriminability between the blinded and unblinded versions of the MTD method? Two factors which could impact

discriminability are suspect-filler similarity and lineup member description-fit, which we measured in the previous chapter. While we had predicted differences between the two MTD methods regarding suspect-filler similarity and lineup member description-fit, these predictions were not confirmed. This suggests that something outside of these two factors is driving these differences in witness discriminability between the two versions of the MTD method.

One possible candidate for this is resultant lineup fairness. While the mock witness task found no real difference between the blinded and unblinded versions of the MTD method, our resultant fairness measures did. Lineups created using the blinded version of the MTD method yielded larger resultant Tredoux's E and effective size values than lineups created using either the unblinded version of the MTD method or the MTS method, with the MTS method yielding the smaller fairness estimates across the board. Additionally, the innocent suspect was less likely to be selected from the blinded MTD lineups than from lineups constructed using the other two methods. This suggests that lineups constructed using the blinded version of the MTD method may better protect innocent suspects from misidentifications. These findings suggest that blinding MTD lineup constructors does impact lineup fairness, when tested with eyewitnesses, rather than mock witnesses.

However, though the results of the ROC analysis show that the method used to create lineups does impact discriminability, the results of the CAC analysis showed that the confidence-accuracy relationship remains relatively unaffected by the lineup construction method. This finding supports research by Semmler and colleagues (2018) which found that suboptimal estimator variables, such as short vs long retention intervals or distance from the perpetrator, had little effect on the confidence-accuracy relationship but did significantly impact discriminability. Though this may seem like a counter-intuitive finding, it is actually predicted by signal detection

theory. This is because when discriminability is impaired, witnesses will fan out their confidence criteria, which in turn allows them to maintain a similar accuracy at each confidence level despite the reduced discriminability. Thus, offering support for signal detection theory as a possible explanation for witnesses to maintain their accuracy even when discriminability has been impaired.

Implications

The results of this study have important implications for the implementation of the MTD method in research and applied settings. Blinding MTD lineup constructors may improve the effectiveness of this method as it led to an increase in empirical discriminability and fairer lineups compared to the unblinded version of the method. Additionally, the results of this study also show that both methods of the MTD method yielded better lineup outcomes compared to the MTS method of lineup construction. Thus, these findings offer support for the use of the MTD method of filler selection over the MTS method. As there are currently no mandated procedures in the United Kingdom or the United States regarding lineup construction, the results of this study have the potential to help improve lineup procedures resulting in better quality lineups. Implications for practice will be discussed in depth in the general discussion.

Limitations and Future Directions

While the findings of this study are promising, there are limitations of the research which should be considered. One limitation of this study is that while the four perpetrators varied from one another regarding age and gender, all four perpetrators in this study were White. The decision to use all White perpetrators was due to the number of plausible fillers within the ethically sourced face database created for the stimulus creation chapter. As the face database

was smaller in size than what law enforcement would use, we had to make the decision to cut down our six initial stimulus sets to four due to a lack of plausible fillers for two of the perpetrators. Thus, future research should be conducted with a more ethnically diverse group of perpetrators as decades of research have shown how the cross-race effect (CRE) impacts various aspects of eyewitness testimony, including the construction and assessment of suspect lineups. Research conducted by Brigham and Ready (1985) investigated the impact the CRE has on lineup composition and fairness. The results showed that lineup constructors were more selective in their filler selections for lineups where the perpetrator was the same race as them (own-race) compared to when they were selecting fillers for other-race lineups. For instance, White lineup constructors took more care to ensure the suspect did not stand out from other lineup members when the suspect was White. However, when the suspect was Black, White lineup constructors took less care to ensure the fillers resembled the suspect (Brigham & Ready, 1985). This in turn could negatively impact the fairness of other-race lineups. Thus, further research should be conducted to investigate how the CRE and filler selection method impacts witness discriminability, confidence, and lineup fairness.

Another potential limitation of this study is the sample size. While we aimed to have a larger sample size, due to time and financial constraints we fell short of our recruitment goal of 1200 participants. This could have led to low statistical power obscuring potential effects, especially in analyses such as CAC where the yield in high confidence identifications was extremely low. For instance, out of 3,358 eyewitness identifications total (including suspect IDs, filler IDs, and non-IDs), there were only 69 correct IDs made from MTS lineups, 81 correct IDs made from blinded MTD lineups, and 56 correct IDs made from unblinded MTD lineups at the highest level of confidence (90-100%). While the results of the CAC analysis showed filler

selection method had an impact on the confidence-accuracy relationship at the lowest levels of confidence, no other significant differences were observed. Thus, because of low yield in identifications made at the highest level of confidence, any potential effects of filler selection method on the confidence-accuracy relationship would be undetectable.

Additionally, there is a need for replication as the mechanism through which blinding MTD lineup constructors affected discriminability remains unclear. The results of this research suggest there may be other factors driving the observed differences in discriminability, as the two factors we initially predicted would impact discriminability (suspect-filler similarity and lineup member description-fit, see Chapter 4) showed no differences between blinded and unblinded MTD lineups. Therefore, further research is needed to investigate how blinding MTD lineup constructors impacts discriminability and to provide further insight into these mechanisms.

Conclusions

In conclusion, the findings of this study offer strong evidence that the MTD method of filler selection, when used properly, can lead to better lineup outcomes and improve lineup quality. Not only did blinded MTD lineups lead to better discriminability compared to MTS lineups, but blinded MTD lineups also led to higher correct ID rates and lower false ID rates compared to MTS lineups. Additionally, the results of this study provide evidence that blinding MTD lineup constructors may improve the effectiveness of the method by increasing empirical discriminability and lineup fairness compared to the unblinded version of the method. While future research should be done to see if the findings from this study can be replicated, this research could have important implications within research and applied settings.

Chapter 6

General Discussion

The purpose of this thesis was to investigate how filler selection method and blinding lineup constructors impacts lineup outcomes such as eyewitness accuracy, confidence, and lineup fairness. To begin, I conducted a systematic review and meta-analysis to gain further insight into the two commonly used filler selection methods, the MTS and MTD methods, as well as how the MTD method has been operationalised in research studies. I found that the research comparing the MTS and MTD methods remains conflicted as to which yields superior lineup outcomes. Additionally, I found that the MTD method of filler selection is rarely implemented using the original methodology developed by Wells and colleagues (1993). This deviation in methodology may contribute to the conflicting findings of previous research.

Next, I investigated whether this change in methodology impacted the effectiveness of the MTD method, and how the process of selecting fillers impacts lineup outcomes. In Study 1, I created the materials that would then be used in subsequent studies. Perpetrator descriptions varying in quality were elicited from eyewitnesses using a free recall questionnaire or a modified version of the Self-Administered Interview (SAI). These descriptions were then used to construct MTD lineups. During the lineup construction phase, participants were provided with: (1) an image of the suspect (MTS condition), (2) a perpetrator description (blinded MTD condition), or (3) an image of the suspect and the matching perpetrator description (unblinded MTD condition). They were then instructed to select fillers who best matched the information provided for a suspect lineup. While lineup constructors in the unblinded MTD condition were instructed to only use the description to base their filler selections on, all but one reported using the image in some capacity during the filler selection process. This admission by unblinded MTD lineup

constructors led me to hypothesise that the unblinded MTD lineups would perform more similarly to the MTS lineups than to blinded MTD lineups.

In Studies 2-4, I evaluated the lineups for fairness, suspect-filler similarity, lineup member description-fit, and eyewitness performance. In Study 2, I tested the lineups for fairness using the mock witness paradigm. Based on existing research, I predicted that the blinded MTD method would yield fairer lineups compared to the MTS method (Jones, 2021). I also predicted that lineups created using the unblinded MTD method would perform more similarly to MTS lineups than to blinded MTD lineups. The first prediction was borne out; the blinded version of the MTD led to fairer lineups than the MTS method, as demonstrated by larger Tredoux's E and effective size estimates. However, the second prediction was not supported; the unblinded MTD method yielded lineup fairness estimates more similar to those of the blinded MTD method than the MTS method.

In Study 3, the lineups were tested for suspect-filler similarity and lineup member fit to the description. As expected, pairwise similarity ratings revealed that lineups constructed using the MTS method led to higher suspect-filler similarity than lineups constructed using the blinded MTD method. Additionally, lineup member fit to the description was rated higher in lineups created using the blinded MTD method than those created using the MTS method. However, contrary to predictions, the unblinded MTD method led to lineups more that were more similar to the blinded MTD method than the MTS method regarding suspect-filler similarity and lineup member description-fit.

Lastly, in Study 4, an eyewitness identification task was used to investigate how filler selection method and blinding MTD lineup constructors impacts witness discriminability, confidence, and resultant lineup fairness. As predicted, blinded MTD lineups yielded better

empirical discriminability than MTS and unblinded MTD lineups. Additionally, resultant lineup fairness measures indicated that blinded MTD lineups were fairer than lineups created using the other two methods.

Below, the key findings of the thesis are discussed. First, I will summarise the results regarding the MTD versus MTS comparison, followed by those of the blinded versus unblinded MTD comparison. Next, practical implications of the research are discussed, along with some of the key limitations of the research and future directions for continued research.

Match-to-Suspect versus Match-to-Description

As discussed in Chapter 1, much research has been done regarding administering and constructing suspect lineups. Recommendations by Wells and colleagues (2020) state that lineups should include only one suspect, alongside a minimum of five ‘quality’ fillers to help reduce the likelihood of eyewitness misidentifications. However, the research remains conflicted as to whether fillers should be selected to resemble the suspect (MTS method) or match the perpetrator description provided by eyewitnesses (MTD method). Both methods have their advantages and disadvantages. An advantage of the MTS method is that it bases filler selection on the appearance of the suspect (Malpass, Tredoux, & McQuiston-Surrett, 2007). Therefore, this method ensures all fillers resemble the suspect, which keeps the suspect from standing out among the other lineup members. This method also follows the guidelines set in the Police and Criminal Evidence Act 1984, Code D, (PACE) which state that fillers should match the suspect on physical characteristics such as race, gender, approximate age, build, and hair colour (Police and Criminal Evidence Act 1984, Code D, 2023). However, researchers have argued that the MTS method of filler selection may yield lineups too high in lineup member similarity which decreases witness discriminability (Luus & Wells, 1991; Wells et al., 1993; Wixted & Mickes,

2014). In efforts to combat the issue of too high lineup member similarity, the MTD method of filler selection was proposed (Luus & Wells, 1991). As the MTD method of filler selection utilises the perpetrator description to select fillers, it ensures that all lineup members match the features mentioned by witnesses but allows for some variation in the features not mentioned (Wells et al., 1993). This variation in features allows witnesses to accurately distinguish between guilty and innocent individuals. However, this method of filler selection is reliant on law enforcement eliciting detailed descriptions from eyewitnesses. Thus, problems with this method arise when witnesses provide vague descriptions or omit “default values”, characteristics which appear obvious to both the witness and officer but need to be confirmed to create an unbiased lineup (Lindsay et al., 1994). Issues aside, when implemented properly, both methods can yield quality lineups which offer protection to innocent suspects.

In Chapter 1, a meta-analysis was conducted to investigate the impact of filler selection method on identification decisions. The results of the meta-analysis found the odds of a correct identification is higher in lineups created using the MTD method compared to lineups created using the MTS method. The meta-analysis also showed that the odds of an innocent suspect identification were also higher for MTD lineups than for MTS lineups. However, this finding was driven by a single, heavily weighted study which may have skewed the results of the impact of filler selection method on innocent suspect identifications. Additionally, filler identifications were more likely to occur in MTS lineups than MTD lineups when the target was present. However, there were no differences between the two methods regarding non-identifications for target present or target absent lineups. The findings of the meta-analysis provided support for the MTD method when the guilty perpetrator was present in the lineup but remained conflicted as to which method leads to better outcomes when the suspect was innocent.

The research conducted for this PhD thesis aimed to provide insight as to which filler selection method, MTS or MTD lead to superior lineup outcomes. The findings of this research offer support for the use of the ‘pure’, or blinded, MTD method of filler selection over the MTS method, supporting one of my main hypotheses. The results of the mock witness paradigm study (Chapter 3) suggest that lineups created using the blinded MTD method yield fairer lineups compared to those created using the MTS method, as measured by Tredoux’s E and effective size. This finding falls in line with existing research which has shown that MTD lineups yield higher Tredoux’s E estimates compared to MTS lineups (Jones, 2021). However, it is worth noting that both effective size and Tredoux’s E measure how many individuals present in the lineup fit the general characteristics of the suspect, or perpetrator description in the case of the mock-witness paradigm (Fitzgerald, Oriet, & Price, 2015; Jones, 2021). Thus, blinded MTD lineup members may match the description provided to mock witnesses better than MTS lineup members, leading to higher Tredoux’s E and effective size estimates. Therefore, to further investigate the influence of filler selection method on lineup fairness, we conducted two follow-up studies investigating lineup member similarity and fit to the description (Chapter 4).

The findings of Study 3a (Chapter 4), revealed that the MTS method led to lineups with higher suspect-filler similarity than lineups created using the blinded MTD method. Conversely, Study 3b (Chapter 4) revealed that the blinded MTD method led to lineups with lineup members who were a better fit to the description than lineup members in MTS lineups. The findings from both studies support existing research which has consistently shown that lineups created using the MTS method are higher in suspect-filler similarity compared to lineups created using the MTD method, but that lineups created using the MTD method tend to have lineup members with better description-fit (Tunnicliff & Clark, 2000; Darling, Valentine, & Memon, 2008; Fitzgerald

et al., 2015). While lineup member similarity and description-fit are important aspects of lineup fairness, measured by perceptual similarity ratings or the mock witness paradigm; there may be other aspects driving lineup fairness estimates that are better measured by resultant lineup size.

In Chapter 5, I calculated resultant lineup fairness measures and again found that lineups created using the blinded MTD method tended to be fairer than lineups created using the MTS method, as measured by Tredoux's E and effective size values. Resultant fairness is considered to be a better measure of lineup fairness as it makes use of the actual eyewitness identification data rather than mock witness data (Quigley-McBride & Wells, 2021). As this method of evaluating lineup fairness uses the eyewitness identification data, it stands to reason it may measure another aspect of lineup fairness – the degree to which the fillers match witnesses' memorial representations of the perpetrator (Quigley-McBride & Wells, 2021; Lee et al., 2025). Lee and colleagues (2025) compared several methods for measuring lineup fairness for concurrent and predictive validity and found little to no correlation between mock witness estimates and resultant fairness estimates. This finding suggests that lineup fairness may be evaluated differently by mock witnesses compared to actual witnesses and may be influenced by eyewitness memory strength for the perpetrator. They also found that resultant lineup fairness estimates were more likely to predict eyewitness performance than mock witness fairness estimates (Lee et al., 2025). The resultant lineup fairness findings of Study 4 (Chapter 5) support those of the mock witness task (Chapter 3) and suggest that lineups created using the blinded MTD method of filler selection yield fairer lineups compared to those created using the MTS method. In addition to impacting lineup fairness, we found that filler selection method had an impact on identification outcomes and empirical discriminability.

The results of the eyewitness identification experiment (Chapter 5) revealed that the

blinded MTD method led to an increase in correct IDs and a reduction in false IDs compared to the MTS method. There were no significant differences in filler IDs or non-IDs between blinded MTD and MTS lineups, regardless of the target being present or absent in the lineup. These findings partially support existing research comparing the MTD method of filler selection to the MTS method which has shown the MTD method leads to higher correct ID rates compared to the MTS method (Wells et al., 1993; Carlson et al., 2019). These studies also found no significant differences between non-ID rates between the two methods, as we did. However, some research has shown no significant differences in false ID rates between the two methods, while others have shown MTD lineups yield higher false ID rates compared to MTS lineups (Wells et al., 1993; Tunnicliff & Clark, 2000; Carlson et al., 2019; Jones, 2021). The results of the meta-analysis (Chapter 1) suggested false IDs were more likely to occur in MTD lineups than MTS lineups. Though this finding should be considered with caution as it was influenced by one heavily weighted study which accounted for 83.49% of the weight, leaving the other five studies to account for the remaining 16.51%. With that said, the results of the identification data (Chapter 5) suggest the opposite of existing literature. We found MTS lineups had a significantly higher rate of false IDs than blinded MTD lineups. While the existing literature is conflicting as to which method yields better lineup outcomes, the findings of this study offer support for the use of the blinded MTD method of filler selection in lineup construction.

These findings are further supported by those of the ROC analysis which showed that empirical discriminability was significantly higher for blinded MTD lineups than for MTS lineups. The findings of the ROC analysis, taken together with those of Study 3a (suspect-filler similarity study), suggest that lineups created using the blinded MTD method allow for more “helpful differences”, or propitious heterogeneity, among lineup members compared to lineups

created using the MTS method (Wells et al., 1993). The findings of this research offer support for the DFD theory, which states that discriminability is increased if eyewitnesses can identify and discount nondiagnostic features between lineup members (Wixted & Mickes, 2014). This then allows witnesses to give more weight to any features that are diagnostic and enables them to distinguish between guilty and innocent individuals. The findings of this research are in line with research conducted by Carlson and colleagues (2019) which found that lineups created using the MTD method led to increased empirical discriminability and propitious heterogeneity compared to MTS lineups.

While the findings of this PhD thesis offer support for the use of the MTD method of filler selection over the MTS method, it is worth noting that this refers to the blinded MTD method specifically. Throughout this thesis, we compared a blinded and unblinded version of the MTD method to the MTS method, as well as to one another, to investigate which filler selection method leads to better lineup outcomes. In Chapter 2, unblinded MTD lineup constructors reported using the suspect's image during the filler selection process. Based on this information and research regarding top-down processing, I hypothesised that lineups created using the unblinded MTD method would perform more similarly to MTS lineups than with blinded MTD lineups. However, the findings from Chapters 3 and 4 did not support this hypothesis.

Study 2 (Chapter 3) tested the lineups for lineup fairness using the mock witness paradigm and found that the unblinded MTD method yielded fairer lineups than the MTS method, as measured by Tredoux's *E* and effective size. This was further supported by the finding that the suspect was significantly less likely to be chosen from lineups created using the unblinded MTD method than those created using the MTS method. These findings offer further support for the use of the MTD method of filler selection over the MTS method, as both blinded

and unblinded MTD lineups led to fairer lineups as tested by mock witnesses.

The next study then investigated whether filler selection method impacted suspect-filler similarity and lineup member description-fit (Chapter 4). Study 3a found that fillers selecting using the MTS method were rated as being more similar to the suspect than fillers selected using the unblinded MTD method. However, there was no difference between the two methods in suspect-filler similarity at the overall lineup level. This finding may be due to the process used to average the individual filler-suspect face pairings to achieve an overall lineup similarity score, therefore minimising any potential differences between the unblinded MTD and MTS method at the lineup level. With that said, Study 3b found differences regarding lineup member description-fit between the unblinded MTD method and MTS method at both the individual trial level (face-description pairings) and the overall lineup level. Lineup members from lineups created using the unblinded MTD method were rated as being a better fit to the description than lineup members from MTS lineups. These findings suggest that even though unblinded MTD lineup constructors had access to the suspect's image, they were still able to select fillers which matched the perpetrator description. With that said, this pattern in results changes in the final empirical chapter of this thesis with the unblinded MTD method performing more similarly to the MTS method on all metrics.

Study 4 (Chapter 5) found no difference between the unblinded MTD method and MTS method regarding identification decisions, regardless of target presence. Both unblinded MTD lineups and MTS lineups yielded similar correct ID and false ID rates, as well as filler IDs and non-IDs for both target present and target absent lineups. This was further supported by the findings of the ROC analysis which found no difference between the two methods in terms of witness discriminability (*see Figure 14*). The unblinded MTD method also performed more

similarly to the MTS method regarding the confidence-accuracy relationship as indicated by the CAC analysis and chi-square analyses comparing both methods at each confidence level (*see Figure 15*). On the final metric, resultant lineup fairness, there was a slight difference in lineup fairness estimates between unblinded MTD lineups and MTS lineups. Lineups created using the MTS method yielded lower Resultant Tredoux's E and Resultant effective size estimates than lineups created using the unblinded MTD method. This finding suggests that the unblinded MTD method may yield fairer lineups than the MTS method, as measured by both mock witnesses and actual eyewitnesses. This change in pattern regarding comparisons between the MTS and unblinded MTD methods may be due to the use of actual eyewitness data in Chapter 5 compared to the use of mock or non-witness data in Chapters 2-4. This suggests that eyewitness memory strength may interact with filler selection method to impact lineup outcomes.

The findings from Chapter 5 may also provide insight as to why the existing literature on filler selection method remains conflicted as to which method leads to better lineup outcomes. Based on the comparisons between the unblinded MTD method and MTS method of filler selection, there is little difference in lineup outcomes between the two methods. However, when looking at the comparisons between the blinded MTD method and the MTS method, the results demonstrate clear support for use of the blinded MTD method over the MTS method. These findings taken together suggest that blinding MTD lineup constructors may in fact improve the effectiveness of the MTD method, at least when it comes to witness accuracy and resultant lineup fairness, offering support for top-down processing theory.

Blinding Match-to-Description Lineup Constructors

The second major research question that underpinned this thesis was whether blinding MTD lineup constructors would lead to superior lineup outcomes. With that said, the findings

throughout Chapters 2-4 were mixed, with the unblinded MTD method performing more similarly to the blinded MTD method on several metrics than to the MTS method. These findings appear to suggest that the blinding of MTD lineup constructors may not be as important as previously believed. However, when it comes to arguably the most important outcome, eyewitness discriminability (Chapter 5), the research suggests that blinding MTD lineup constructors does matter.

In Chapter 5 (Study 4), there was clear evidence of superiority for blinded MTD lineups over both unblinded MTD lineups and MTS lineups. The results of the identification data showed that lineups created using the blinded version of the MTD method led to significantly more correct IDs and fewer false IDs compared to lineups created using the unblinded version of the method. This is further supported by the findings of the ROC analysis, which found blinding MTD lineup constructors led to lineups with higher empirical discriminability compared to MTD lineups where the constructors were not blinded. The findings of the ROC analysis showed no significant differences in discriminability between MTS and unblinded MTD lineups, this was demonstrated by the ROC curves for MTS and unblinded MTD lineups laying nearly on top of one another (*see Figure 14*).

While it is clear blinding MTD lineup constructors does have an impact on witness accuracy for MTD lineups, it is unclear what is driving these differences in discriminability. On metrics thought to impact discriminability, such as mock witness lineup fairness (Chapter 3), suspect-filler similarity ratings (Chapter 4), or lineup member description-fit (Chapter 4), there were few significant differences between the two MTD methods. Research on top-down processing explains how prior knowledge and experience influences perception and decision-making (Firestone & Scholl, 2014; Kinchla & Wolfe, 1979). By providing unblinded MTD

lineup constructors with an image of the suspect, I predicted they would be unable to disregard the knowledge gained from the image during the filler selection process and thus, they would base their filler selections, at least in part, on the appearance of the suspect. This was supported by the filler selection process information provided by unblinded MTD lineup constructors in Study 1b (Chapter 2). Of the eight unblinded MTD constructors, seven reported using the image of the suspect to assist in the selection of fillers. Despite being instructed to use the perpetrator description to select fillers, four of the unblinded MTD constructors reported only using the image and disregarded the description provided, while an additional three reported using the image in combination with the description. This finding taken with the existing research on top-down processing informed my hypothesis that unblinded MTD lineups would perform more similarly to MTS lineups on all metrics than to blinded MTD lineups.

While research has shown that lineup fairness impacts discriminability, the findings from Study 2 (Chapter 3) found no difference in mock witness lineup fairness estimates between the two MTD methods (Colloff et al., 2016). This finding was surprising given there was a significant difference found between the two MTD methods regarding the likelihood of the suspect being chosen from the lineup by mock witnesses. The suspect was more likely to be chosen from lineups created using the unblinded MTD method than from blinded MTD lineups. This suggests that blinding may not impact lineup fairness as measured by lineup size estimates (i.e. Tredoux's E and effective size) but may impact another aspect of lineup fairness – lineup bias. This finding also suggests that while unblinded MTD constructors were using the image in some capacity to select fillers, they were also successful in selecting fillers who matched the perpetrator description. However, it is worth noting that while there was no difference in the mock witness fairness estimates between the two MTD methods (Chapter 3), there was a

difference between the two MTD methods regarding resultant lineup fairness estimates (Chapter 5). Resultant Tredoux's E and resultant effective size estimates were higher for blinded MTD lineups than for unblinded MTD lineups, suggesting blinding MTD lineup constructors may improve lineup fairness when measured by resultant lineup fairness estimates. This finding supports research by Lee and colleagues (2025) which found resultant lineup fairness estimates had weak or non-significant correlations with those from the mock witness paradigm, suggesting poor concurrent validity. As discussed in Chapter 3, this may be due to the mock witness paradigm only measuring one indicator of lineup fairness – the degree to which lineup members match a perpetrator description. Resultant lineup fairness, on the other hand, may take into account additional aspects of fairness such as suspect-filler similarity and an eyewitness's memory strength for the perpetrator.

With that said, research has shown that suspect-filler similarity may impact lineup fairness and discriminability (Carlson et al., 2008; Colloff et al., 2016). Study 3a and 3b (Chapter 4) investigated whether blinding MTD lineup constructors impacts the effectiveness of the method in terms of suspect-filler similarity and lineup member fit to the description. Contrary to predictions, the unblinded version of the MTD method yielded suspect-filler similarity and lineup member description-fit ratings more comparable to the blinded MTD method than to the MTS method. In Study 3a, there was no difference in suspect-filler similarity between the two MTD methods at either the individual face-pair level or at the overall lineup level. However, in Study 3b, there was a significant difference between the two MTD methods in lineup member description-fit at the individual trial level (i.e. for individual face-description pairs). Lineup members selected by unblinded MTD lineup constructors were rated as fitting the description better than those selected by blinded MTD lineup constructors. With that said, this difference in

description-fit ratings between the two MTD methods was only seen at the individual trial level and not at the overall lineup level. These findings suggest that while unblinded MTD lineup constructors had access to the suspect's image, they were still able to select lineup members who match the description. These findings also imply that blinding MTD lineup constructors may not have an impact on suspect-filler similarity or lineup member description-fit. Thus, the results of Study 3a and 3b indicate that something other than lineup member similarity and description-fit are driving the differences in discriminability that emerged in Study 4 (Chapter 5).

There are several factors which could potentially interact with blinding MTD lineup constructors to impact discriminability, including description quality, face database size, and eyewitness memory strength. For instance, blinding MTD lineup constructors would have less of an impact when the perpetrator description is extremely detailed compared to when the description is vague. Imagine selecting fillers for a lineup using the following description, "Perpetrator is a White male, early to mid 20s, approximately 5'10", athletic build, has a round face with freckles across his nose and cheeks, with short curly auburn hair, thick eyebrows, and blue almond shaped eyes." With a description this detailed, there will likely only be so many individuals within the face database who match this description. Thus, blinding lineup constructors using the MTD method would have a limited impact on the selection of fillers, because the pool of possible fillers would be small. However, if the description is vague and lacking those extra details, having access to an image of the suspect may lead to a greater difference in the fillers selected compared to fillers selected using only the description; this may occur because there is a larger pool of potential fillers to choose between. Therefore, it is possible description quality may interact with the blinding of MTD lineup constructors to impact discriminability.

Though we initially intended to investigate the impact description quality had on MTD lineups by creating low- and high-quality descriptions that varied in the number of descriptors provided, we were unable to as ultimately the differences between the two description conditions were minimal. Not only was there little difference in number of descriptors provided between the two description conditions, including description quality as an independent variable throughout this research would have led to complex experimental designs that would have required larger sample sizes than we were able to achieve. Therefore, we made the decision to use description quality as a random variable for generalisability purposes as in real-world scenarios, police do not always have detailed perpetrator descriptions available. Thus, both blinded and unblinded MTD lineup constructors had access to perpetrator descriptions varying in quality, which may have led to some variability within the MTD lineups created.

In addition to affecting discriminability, blinding and description quality may interact to impact lineup fairness. Research by Mansour and colleagues (2017) has shown the mock witness paradigm is sensitive to changes in description quality. While the study by Mansour and colleagues (2017) used lineups created using a variation of the MTS method, the results of the study show that changes in the quality of the descriptions provided to mock witnesses impacted lineup fairness. This was done by providing mock witnesses with either a verbatim ‘individual’ description from an eyewitness, a modal or ‘consensus’ description containing only features mentioned by half of the eyewitnesses, or a comprehensive description containing all non-contradictory details mentioned by eyewitnesses. Comprehensive descriptions included the most detail, with consensus descriptions being the least detailed. They found that lineups appeared to be fairer when mock witnesses read an individual description or a comprehensive description, compared to reading a consensus description. This suggests that providing mock witnesses with

more detailed descriptions may increase lineup fairness.

Therefore, a potential explanation for the lack of significant differences in mock witness fairness estimates between the blinded and unblinded MTD conditions may be due to our use of consensus, or modal descriptions in this research. The use of consensus descriptions may have led to our descriptions containing fewer descriptors than if we had used individual witness descriptions or created comprehensive descriptions. As discussed in Chapter 2, the number of descriptors reported by eyewitnesses in high-quality descriptions ranged from 5 to 7.33 descriptors, while the low-quality descriptions ranged from 3.75 to 5.25 descriptors. This amount of detail is consistent with existing research which has shown witnesses report on average 3.9 to 9.7 physical descriptors when describing a perpetrator (Lindsay et al., 1994; Dalton, 2021). While both low- and high-quality descriptions were provided to blinded and unblinded MTD lineup constructors to select fillers, only high-quality descriptions were provided to mock witnesses to evaluate the lineups. Thus, even though we used the high-quality descriptions to evaluate the lineups for fairness, they may not have contained enough detail to allow the mock witness task to detect differences in lineup fairness between the two MTD methods.

Another factor which could interact with the blinding of MTD lineup constructors and description quality to impact discriminability is the size of the face database used to select fillers. Research by Bergold and Heaton (2018) demonstrated face database size impacts suspect-filler similarity and discriminability; such that as database size increases, so does suspect-filler similarity, while negatively impacting discriminability. With that said, it stands to reason a larger face database would yield a larger number of plausible fillers who could be selected. This increase in plausible fillers could lead to a greater difference in the fillers selected by blinded and unblinded MTD lineup constructors, especially if the perpetrator description is rather vague and

unblinded MTD lineup constructors are using their knowledge of the suspect's appearance to aid their filler selections. This interaction between face database size and blinding is important as face databases used by law enforcement in the UK contain thousands of faces (i.e. the Video Identification Parade Electronic Recording, also known as VIPER) which can lead to greater differences in the fillers selected by blinded and unblinded MTD lineup constructors. Therefore, this divergence in the choice of fillers may in turn impact lineup outcomes such as discriminability and fairness. However, it is worth noting that these are just speculations about possible factors which could contribute to the findings of this research. Further research should be conducted to determine how blinding MTD lineup constructors impacts the effectiveness of the method.

Implications

The findings of this research have important implications for policy and procedure regarding lineup construction both locally and globally. Lineup procedures not only vary greatly from country to country but also vary between different jurisdictions within the same country (Fitzgerald, Rubínová, & Juncu, 2021; Wells et al., 2020). A meta-analysis by Fitzgerald, Rubínová, and Juncu (2021) found the MTS method of filler selection is recommended for use by nearly all policing guidelines (89%), with only 17% of policing guidelines recommending the use of the MTD method for lineup construction. It is worth noting several of the guidelines that recommend using the MTD method also recommend matching fillers to the suspect's appearance, which is why these figures sum to greater than 100%. However, the results of this thesis offer strong support for selecting fillers who match the description of the perpetrator over fillers who match the suspect's appearance.

Implementing the MTD method would be a simple and cost-effective change in

procedure in the UK. The current guidelines set in the Police and Criminal Evidence Act 1984, Code D, (PACE), state all fillers should generally resemble the suspect, but there is no additional guidance regarding the filler selection process. While the findings of this thesis suggest that MTS lineups yield higher suspect-filler similarity, MTD lineups yielded fairer lineups when measured by both the mock witness paradigm and resultant lineup fairness. This suggests that no single lineup member, including the suspect, stood out from the others in MTD lineups. Additionally, the purpose of a video identification as stated in PACE, Code D, Annex A, Note A1 “is to test the eye-witness’ ability to distinguish the suspect from others and it would not be a fair test if all the images shown were identical or extremely similar to each other” (Police and Criminal Evidence Act 1984, Code D, 2023, p. 39). Thus, the PACE guidelines suggest some variation is needed among lineup members to be considered fair. This offers support for the implementation of the MTD method in the construction of lineups in the UK.

The findings of this research also offered support for the blinded version of the MTD method over the unblinded MTD method regarding witness discriminability. As the PACE guidelines already dictate the identification officer should be “an officer not below inspector rank who is not involved with the investigation”, it would be fairly simple to implement a blinded version of the MTD method in practice (Police and Criminal Evidence Act 1984, Code D, 2023, p. 12). Thus, as long as the identification officer is kept separate from the suspect and any images of the suspect, it would be possible to implement the blinded version of the MTD method under the existing PACE Codes. However, once all fillers have been selected it would be beneficial for identification officers to view the completed lineup to ensure all lineup members match the description and that no individual, including the suspect, stands out from the others. Additionally, there may be instances where use of the MTS method or a combination of the two

methods would be preferable in the construction of lineups.

Researchers have argued use of the MTD method would be impractical in instances where the suspect does not match the perpetrator description, the perpetrator description is vague or lacking necessary detail, and when the suspect has distinctive features (Luus & Wells, 1991; Lindsay et al., 1994; Wells et al., 2020). In instances where an individual becomes a suspect based on evidence other than their appearance, it is possible for there to be discrepancies between the description and suspect's appearance. In those cases, it would be more appropriate to match the fillers to the features of the suspect that deviate from the description to prevent them from standing out in the lineup. Lindsay and colleagues (1994) argued that vague perpetrator descriptions are problematic for the MTD method of filler selection as they can lead to the creation of biased lineups. In instances where only a vague description has been elicited from witnesses, it would be beneficial to use a blended methodology (Wells et al., 2020). Such that all fillers match the minimal details mentioned in the description and then match fillers to the suspect in necessary details which were omitted from the description (Lindsay et al., 1994; Wells et al., 2020). Additionally, when the suspect has distinctive features (i.e. scars, tattoos, or birth marks) steps should be taken to conceal or replicate the feature among other lineup members to ensure the suspect does not stand out in the lineup (Colloff et al., 2016). This recommendation is in line with the guidelines laid out in PACE, Code D (Police and Criminal Evidence Act 1984, Code D, 2023). When this occurs, it would be prudent to use the MTS method or a combination of the two methods to select fillers rather than the blinded MTD method. However, though there may be some cases in which the MTS method may be required, in most situations the use of the blinded MTD method would be appropriate and easily implemented under the PACE guidelines.

As eyewitness testimony remains a persuasive form of evidence, it is vital that best

practices be used to reduce eyewitness misidentifications. Data from the National Registry of Exonerations (2025) has shown that eyewitness misidentification played a role in 27% of all wrongful convictions in the United States. Data from the Miscarriages of Justice Registry (2025) has found similar numbers in the United Kingdom, with eyewitness evidence playing a role in nearly 25% of wrongful convictions. Therefore, it is important for researchers and policymakers to come together to find a way to bridge the gap between research and practice to ensure police are using lineup procedures with the best possible outcomes.

Limitations and Future Directions

There are a few limitations to this research which could have underestimated the effect of blinding on MTD lineup constructors. One limitation of this study is the size of the face database that lineup constructors used for their filler selections. As we wanted to only use ethical databases in which individuals had given permission for their faces to be used in research, the database for this project was smaller than what would be used by police as it had only 667 faces. Research by Bergold and Heaton (2018) investigated the impact face database size has on identification accuracy by creating small (5,000 photos), medium (25,000 photos), and large (125,000 photos) face databases used to select fillers. They found that as face database size increased, so did suspect-filler similarity. Additionally, they found as face database size increased, identification accuracy and confidence decreased. These findings suggest face database size may play an important role in the construction of fair suspect lineups by impacting suspect-filler similarity. As the database used in this research was much smaller than even those used by Bergold and Heaton (2018), it stands to reason it could have led to a decrease in effect size. As previously discussed, it is plausible the use of a smaller face database led to a smaller number of plausible fillers, which may have underestimated the effect of blinding on the

effectiveness of the MTD method.

Despite using a relatively small face database, there was variation in the fillers selected for all the lineups. In Chapter 2, the number of unique fillers showed the same five fillers were not selected for each lineup within an individual stimulus set. Additionally, the results of the Pearson correlation showed that there was little overlap in fillers between lineups created using the blinded MTD method and the MTS method. However, the Pearson correlation did reveal a moderate correlation in fillers selected between the two MTD methods. This finding suggests there was more overlap in filler selection between the two MTD methods and may be the result of using the smaller face database compared to a larger, real-world database. Thus, this overlap in fillers due to database size may have underestimated the effect of blinding on lineup constructors using the MTD method of filler selection. Future research should use a larger face database, such as VIPER, to further investigate how filler selection method and blinding MTD lineup constructors impacts lineup outcomes.

Another way in which having a smaller face database may have impacted this research is by limiting the diversity among the stimulus sets. Due to the smaller face database, two stimulus sets were dropped prior to the lineup construction phase as there were not enough individuals who resembled the suspect. While I ended up with four different perpetrators ranging in age and gender, all perpetrators were White. This lack of diversity in race of the perpetrators may have impacted the results of this research as decades of research have shown how the cross-race effect impacts various aspects of eyewitness testimony, including the construction and assessment of suspect lineups. Research on the cross-race effect has shown that lineup constructors are more selective about fillers to include in lineups where the suspect is the same race as them (own-race), compared to when the suspect is of another race (other-race) (Brigham & Ready, 1985).

For instance, White lineup constructors took more time and care to ensure the suspect did not stand out from other lineup members when the suspect was White. However, when the suspect was Black, White lineup constructors spent less time and took less care to ensure the fillers resembled the suspect (Brigham & Ready, 1985). This difference in selectivity may in turn negatively impact the fairness of other-race lineups. Thus, Brigham and colleagues (2007) recommended that lineup constructors should be of the same race as the suspect they are constructing the lineup for. As not all lineup constructors in this research were White, future research utilising a more diverse stimulus set may be able to minimise possible cross-race effects in the construction and assessment of lineups.

Not only may race play an important factor in the construction and assessment of lineups, but it may also impact the quality of descriptions elicited from eyewitnesses. Research by Ellis, Deregowski, and Shepherd (1975) showed race effects between Black and White participants when asked to describe own-race and other-race faces. They found Black participants mentioned more features than White participants regardless of the race of the individual they were describing. They also reported a difference in the types of features mentioned by Black and White participants. For example, Black participants were more likely to mention features regarding hair position and styling, eye size and whites of the eyes, eyebrows, ears, and chin (Ellis et al., 1975; Brigham et al., 2007). White participants, on the other hand, were more likely to report features such as eye colour as well as hair colour and texture (Ellis et al., 1975; Brigham et al., 2007). Not only were there differences in the types of features mentioned, but White participants were also more likely to provide a vague, general description of Black faces which could be problematic for law enforcement officers tasked with eliciting perpetrator descriptions (Ellis et al., 1975). These findings suggest that race could negatively impact the

quality of the perpetrator description elicited by law enforcement if the perpetrator is of a different race than the witness. As the success of the MTD method is reliant on quality, detailed perpetrator descriptions, it is important race of the witness and perpetrator are considered by officers when eliciting descriptions for an investigation.

Another limitation of this research is the dropping of the high- and low-quality description manipulation, specifically for the mock witness and description-fit studies. For both the mock witness task and the lineup member description-fit studies, only the high-quality perpetrator descriptions were provided to participants. While there was not a large difference in the number of descriptors provided between the high-quality ($M = 6.29$, $SE = 0.44$) and low-quality ($M = 4.37$, $SE = 0.29$), the results of the hierarchical regression were significant (Chapter 2), suggesting I was successful in manipulating description quality. As the difference in descriptors between the low- and high-quality description conditions was only about 2 descriptors, it would be more realistic to use the best quality description available to evaluate the lineups. As such, the use of only high-quality perpetrator descriptions may have impacted the lineup member description-fit ratings, as well as the mock witness fairness estimates. As discussed in Chapter 3, the mock witness task appears to only measure one facet of lineup fairness – the degree to which lineup members match the perpetrator description. Therefore, it is possible that using only the high-quality descriptions led to both MTD conditions yielding fairer lineup estimates compared to MTS lineups. This is because it is more likely for all lineup members in MTD lineups to match the features mentioned in the description than for MTS lineup members who were selected to resemble the suspect. The same could be said for the study on lineup member description-fit. As research by Mansour and colleagues (2017) has shown the mock witness task to be sensitive to differences in methodology, such as description quality,

future research should be conducted to investigate how description quality impacts mock witness fairness and lineup member description-fit.

While this research has its limitations, it sets the groundwork for future research to further investigate the impact that blinding may have on the effectiveness of the MTD method. Future research should investigate the interaction between blinding and description quality, as well as database size. This would allow researchers to determine the limits of this method. For example, is there a point at which too much detail in the description leads to the blinded MTD method yielding similar suspect-filler similarity ratings to MTS lineups? On the other hand, what is the minimum amount of detail a description can have before it leads to blinded MTD constructors creating biased lineups? Finding these limits of the method could have important implications as research on the cross-race effect has shown that the type of descriptors provided differs based on the race of the eyewitness and perpetrator. Additionally, database size may interact with blinding and description quality in an important way. Research has shown that as database size increases, so does suspect-filler similarity (Bergold & Heaton, 2018). However, Bergold and Heaton (2018) used a multiple step approach to construct the lineups used in their study. First, they used facial algorithm program to purposely select the 12 most similar faces within the face database. They then had a former homicide detective select the five fillers they believed should go into the lineup alongside the suspect. Thus, it is possible that the MTD method could mitigate the risks posed by very large face databases. Therefore, future research should investigate whether face database size and filler selection method, including a blinded version of the MTD method, has an impact on suspect-filler similarity and lineup outcomes.

Conclusion

This thesis aimed to determine whether the MTS or MTD method of filler selection

yields better quality lineups and improves witness accuracy. It also aimed to investigate whether blinding MTD lineup constructors to the suspect's appearance improves the effectiveness of the method. Across every measure, we found that the blinded MTD method produced superior lineup outcomes compared to the MTS method, offering support for the use of the MTD method of filler selection. Additionally, the findings of this research suggest blinding MTD lineup constructors may improve the effectiveness of the MTD method as demonstrated by better witness discriminability. However, the driving factors behind this increase in discriminability remain unclear. Thus, further research should be conducted to determine how blinding lineup constructors using the MTD method improves the method's effectiveness. As data from exoneration registries in the US and UK have shown eyewitness misidentifications play a role in nearly a quarter of all wrongful convictions, it is important that researchers and policymakers come together to improve the policies and procedures surrounding suspect lineups.

References

- Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior research methods*, 52(1), 388-407.
- Baldassari, Moore, Hyman, Hope, Mah, Lindsay, Mansour, Saraiva, Horry, Rath, Kelly, Jones, Vale, Lawson, Pedretti, Palma, Cruz, Quarenta, Cruyssen, Mileva, Allen, Jeye, & Wiechert. (2023). The effect of pre-event instructions on eyewitness identification. *Cognitive Research: Principles and Implications*, 8(16). DOI:10.1186/s41235-023-00471-4.
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of statistical software*, 67, 1-48.
- Bellefontaine, S. P., & Lee, C. M. (2014). Between black and white: Examining grey literature in meta-analyses of psychological research. *Journal of Child and Family Studies*, 23(8), 1378-1388.
- Bergold, A. N., & Heaton, P. (2018). Does filler database size influence identification accuracy?. *Law and Human Behavior*, 42(3), 227.
- Bornstein, Deffenbacher, Penrod, & McGorty. (2012). Effects of exposure time and cognitive operations on facial identification accuracy: a meta-analysis of two variables associated with initial memory strength. *Psychology, Crime & Law*, 18(5), 473-490. DOI: 10.1080/1068316X.2010.508458.
- Brehm, L., & Alday, P. M. (2022). Contrast coding choices in a decade of mixed models. *Journal of Memory and Language*, 125, 104334.

- 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.
- Breznitz, Z. (2002). Asynchrony of visual-orthographic and auditory-phonological word recognition processes: An underlying factor in dyslexia. *Reading and Writing*, 15(1), 15-42.
- Brigham, Bennett, Meissner, & Mitchell. (2007). The influence of race on eyewitness memory. In *The Handbook of Eyewitness Psychology: Volume II* (pp. 271-296). Psychology Press.
- Brigham, J. C., & Ready, D. J. (1985). Own-race bias in lineup construction. *Law and Human Behavior*, 9(4), 415-424.
- Brown, Lloyd-Jones, & Robinson. (2008). Eliciting person descriptions from eyewitnesses: A survey of police perceptions of eyewitness performance and reported use of interview techniques. *European Journal of Cognitive Psychology*, 20(3), 529–560.
<https://doi.org/10.1080/09541440701728474>
- Burns, E. J. (2024). Improving the DSM-5 approach to cognitive impairment: Developmental prosopagnosia reveals the need for tailored diagnoses. *Behavior Research Methods*, 56(7), 7872-7891.
- Burns, E. J., Gaunt, E., Kidane, B., Hunter, L., & Pulford, J. (2023). A new approach to diagnosing and researching developmental prosopagnosia: Excluded cases are impaired too. *Behavior Research Methods*, 55(8), 4291-4314.

- Carlson, C. A., Jones, A. R., Whittington, J. E., Lockamy, R. F., Carlson, M. A., & Wooten, A. R. (2019). Lineup fairness: Propitious heterogeneity and the diagnostic feature-detection hypothesis. *Cognitive research: principles and implications*, 4(1), 20.
- Carlson, Gronlund, & Clark. (2008). Lineup composition, suspect position, and the sequential lineup advantage. *Journal of Experimental Psychology: Applied*, 14(2), 118-128.
- Carlson, Young, Weatherford, Carlson, Bednarz, & Jones. (2016). The influence of perpetrator exposure time and weapon presence/timing on eyewitness confidence and accuracy. *Applied Cognitive Psychology*, 30(6), 898-910. <https://doi.org/10.1002/acp.3275>.
- Charman, S. D., & Quiroz, V. (2016). Blind sequential lineup administration reduces both false identifications and confidence in those false identifications. *Law and Human Behavior*, 40(5), 477.
- 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.
- 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.
- Dalton, G. (2021). 'He looks like your typical average guy': Eliciting person descriptions during frontline interviews. [Doctoral dissertation, The University of Portsmouth]. https://pure.port.ac.uk/ws/portalfiles/portal/53405951/PhD_Final_Thesis_.pdf.
- Darling, S., Valentine, T., & Memon, A. (2008). Selection of lineup foils in operational contexts. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, 22(2), 159-169.

DeBruine & Jones. (2017). Face Research Lab London Set. figshare. Dataset.

<https://doi.org/10.6084/m9.figshare.5047666.v5>

Deffenbacher, Bornstein, McGorty, & Penrod. (2008). Forgetting the once-seen face: Estimating the strength of an eyewitness's memory representation. *Journal of Experimental Psychology: Applied*, 14(2), 139-150. doi 10.1037/1076-898X.14.2.139.

Demarchi & Py. (2009). A method to enhance person description: A field study. *Handbook of psychology of investigative interviewing: Current developments and future directions*, 241-256.

Doob & Kirshenbaum. (1973). Bias in police lineups—Partial remembering. *Journal of Police Science and Administration*, 1(3), 287–293.

Duchaine & Nakayama. (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.

Duval, S., & Tweedie, R. (2000). Trim and fill: a simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, 56(2), 455-463.

Ebner, Riediger, & Lindenberger. (2010). FACES—A database of facial expressions in young, middle-aged, and older women and men: Development and validation. *Behavior Research Methods*, 42(1), 351-362.

Ellis, H. D., Deregowski, J. B., & Shepherd, J. W. (1975). Descriptions of white and black faces by white and black subjects (1). *International Journal of Psychology*, 10(2), 119-123.

Evelo, A., Lee, J., Modjadidi, K., & Penrod, S. (2018, July). The role of lineup bias in witness accuracy, the confidence-accuracy relationship and the courtroom value of witness

- confidence. In *Annual Conference of the European Association of Psychology and Law, Turku, Finland*.
- Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39, 175-191.
- Firestone, C., & Scholl, B. J. (2014). “Top-down” effects where none should be found: The El Greco fallacy in perception research. *Psychological science*, 25(1), 38-46.
- Fitzgerald, R., Oriet, C., & Price, H. (2015). Suspect filler similarity in eyewitness lineups: A literature review and a novel methodology. *Law and Human Behavior*, 39(1), 62.
- Fitzgerald, R. J., Price, H. L., Oriet, C., & Charman, S. D. (2013). The effect of suspect-filler similarity on eyewitness identification decisions: A meta-analysis. *Psychology, Public Policy, and Law*, 19(2), 151.
- 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.
- Fitzgerald, R. J., Whiting, B. F., Therrien, N. M., & Price, H. L. (2014). Lineup member similarity effects on children's eyewitness identification. *Applied Cognitive Psychology*, 28(3), 409-418.
- Gabbert & Brown. (2015). Interviewing for face identification. *Forensic facial identification: Theory and practice of identification from eyewitnesses, composites and CCTV*, 15-41.
- Gabbert, Hope, & Fisher. (2009). Protecting eyewitness evidence: Examining the efficacy of the self-administered interview tool. *Law and Human Behavior*, 33(4), 298- 307.
- <https://doi.org/10.1007/s10979-008-9146-8>.

- Gabbert, Hope, Fisher, & Jamieson. (2012). Protecting against misleading post-event information with a self-administered interview. *Applied Cognitive Psychology*, 26(4), 568-575. <https://doi.org/10.1002/acp.2828>.
- Gabbert, Tamonyte, Apps, Caso, Woolnough, Hope, Handscomb, & Waterworth. (2020). Examining the efficacy of a self-administered report form in missing persons investigations. *Legal and Criminological Psychology*, 25, 1-16.
- Geiselman, Fisher, Firstenberg, Hutton, Sullivan, Avetissian, & Prosk. (1984). Enhancement of eyewitness memory: An empirical evaluation of the cognitive interview. *Journal of Police Science and Administration*, 12, 74 – 80.
- Gronlund, S. D., & Neuschatz, J. S. (2014). Eyewitness identification discriminability: ROC analysis versus logistic regression. *Journal of Applied Research in Memory and Cognition*, 3(2), 54-57.
- Gronlund, S., Wixted, J., & Mickes, L. (2014). Evaluating eyewitness identification procedures using receiver operating characteristic analysis. *Current Directions in Psychological Science*, 23(1), 3-10.
- Haw & Fisher. (2004). Effects of administrator-witness contact on eyewitness identification accuracy. *Journal of Applied Psychology*, 89(6), 1106
- Home Office. (2023). *PACE Code D: Code of Practice for the identification of persons by police officers*. <https://assets.publishing.service.gov.uk/media/65816ec7fc07f300128d4433/PA-CE+Code+D+2023.pdf>.
- Hope, Gabbert, & Fisher. (2011). From laboratory to the street: Capturing witness memory using the Self-Administered Interview. *Legal and Criminological Psychology*, 16, 211-226. <https://doi.org/10.1111/j.2044-8333.2011.02015.x>.

- Horry, Hughes, Hope, & Gabbert. (2022). Capturing best evidence from witnesses to serious road traffic collisions: A field trial of the Self-Administered Interview for Road Traffic Collisions. Final Report for the Road Safety Trust.
- Horry, R., Fitzgerald, R. J., & Mansour, J. K. (2021). “Only your first yes will count”: The impact of prelineup instructions on sequential lineup decisions. *Journal of Experimental Psychology: Applied*, 27(1), 170–186. <https://doi.org/10.1037/xap0000337>
- The Innocent Project. (2024). Eyewitness identification. <https://innocenceproject.org/eyewitness-misidentification/>.
- JASP Team (2024). JASP (Version 0.18.3)[Computer software].
- Jones, A. (2021). The effect of filler selection methods on lineup fairness and eyewitness identification. (Doctoral Dissertation, Texas A&M University-Commerce).
- Juslin, P., Olsson, N., & Winman, A. (1996). Calibration and diagnosticity of confidence in eyewitness identification: Comments on what can be inferred from the low confidence–accuracy correlation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(5), 1304.
- Kask & Bull. (2009). The effects of different presentation methods on multi-ethnicity face recognition. *Psychology, Crime & Law*, 15(1), 73-89.
- Kinchla, R. A., & Wolfe, J. M. (1979). The order of visual processing: “Top-down,” “bottom-up,” or “middle-out”. *Perception & psychophysics*, 25(3), 225-231.
- Koehnken, G., Malpass, R. S., & Wogalter, M. S. (1996). Logical foundations for the identification of individuals. *Psychological Issues in Eyewitness Identification*, 205.
- Kovera, M. B., & Evelo, A. J. (2017). The case for double-blind lineup administration. *Psychology, Public Policy, and Law*, 23(4), 421.

- Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. (2017). lmerTest package: tests in linear mixed effects models. *Journal of statistical software*, 82, 1-26.
- Lee, Mansour, & Penrod. (2022). Validity of mock-witness measures for assessing lineup fairness, *Psychology, Crime & Law*, 28(3), 215-245, DOI:10.1080/1068316X.2021.1905811.
- Lee, Mansour, & Penrod. (2025) How to measure lineup fairness: concurrent and predictive validity of lineup-fairness measures. *Psychology, Crime & Law*, 31(6), 666-690, DOI: 10.1080/1068316X.2024.2307358.
- 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 and Human Behavior*, 18(5), 527-541.
- Lucas, Brewer, & Palmer. (2021). Eyewitness identification: The complex issue of suspect-filler similarity. *Psychology, Public Policy, and Law*, 27(2), 151.
- Luus, C. E., & Wells, G. L. (1991). Eyewitness identification and the selection of distracters for lineups. *Law and Human Behavior*, 15(1), 43-57.
- Ma, Correll, & Wittenbrink. (2015). The Chicago Face Database: A free stimulus set of faces and norming data. <https://doi.org/10.3758/BRM.42.1.351>.
- MacLean, C. L. (2019). The Self-Administered Witness Interview Tool (SAW-IT): Enhancing witness recall of workplace incidents. *Applied Cognitive Psychology*, 33, 1212-1223. DOI: 10.1002/acp.3568.
- Malpass, R. S. (1981). Effective size and defendant bias in eyewitness identification lineups. *Law and Human Behavior*, 5, 299–309. doi: 10.1007/BF01044945.

- Malpass, R. S., & Lindsay, R.C.L. (1999). Measuring line-up fairness. *Applied Cognitive Psychology*, 13, S1–S7. doi:10.1002/(SICI)1099-0720(199911)13:1+<S1::AIDACP678>3.0.CO;2-9
- Malpass, R. S., Tredoux, C. G., & McQuiston-Surrett, D. (2007). Lineup construction and lineup fairness. In R. C. L. Lindsay, D. F. Ross, J. D. Read, & M. P. Toglia (Eds.), *The handbook of eyewitness psychology*, Vol. 2. Memory for people (pp. 155–178). Lawrence Erlbaum Associates Publishers.
- Mansour, J. K. (2020). The confidence-accuracy relationship using scale versus other methods of assessing confidence. *Journal of Applied Research in Memory and Cognition*, 9(2), 215–231. <https://doi.org/10.1037/h0101846>
- Mansour, J. K., Beaudry, J. L., Kalmet, N., Bertrand, M. I., & Lindsay, R. C. L. (2017). Evaluating lineup fairness: Variations across methods and measures. *Law and Human Behavior*, 41(1), 103–115. <https://doi.org/10.1037/lhb0000203>.
- Meissner, C. A., & Brigham, J. C. (2001). Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychology, Public Policy, and Law*, 7(1), 3.
- Meissner, Sporer, & Schooler. (2014). Person descriptions as eyewitness evidence. In *Handbook of Eyewitness Psychology 2 Volume Set* (pp. 3-34). Routledge.
- Memon, Hope, & Bull. (2003). Exposure duration: Effects on eyewitness accuracy and confidence. *British Journal of Psychology*, 94, 339-354.
- Menne, N. M., Winter, K., Bell, R., & Buchner, A. (2023). Measuring lineup fairness from eyewitness identification data using a multinomial processing tree model. *Scientific Reports*, 13(1), 6290.

- Milne, R., & Bull, R. (2006). Interviewing victims of crime, including children and people with intellectual disabilities. *Practical psychology for forensic investigations and prosecutions*, 7, 24.
- Miscarriages of Justice Registry. (2025). Miscarriages of Justice Registry.
<https://evidencebasedjustice.exeter.ac.uk/miscarriages-of-justice-registry/the-cases/overview-graph/>
- Molinaro, Charman, & Wylie. (2021). Pre-identification confidence is related to eyewitness lineup identification accuracy across heterogeneous encoding conditions. *Law and Human Behavior*, 45(6), 524–541. <https://doi.org/10.1037/lhb0000452>.
- National Registry of Exonerations. (2025). The National Registry of Exonerations.
https://exonerationregistry.org/cases?f%5B0%5D=n_pre_1989%3A0.
- Nicholls, Churches, & Loetscher. (2018). Perception of an ambiguous figure is affected by own-age social biases. *Scientific reports*, 8(1), 12661.
- Palmer, Brewer, Weber, & Nagesh. (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. DOI: 10.1037/a0031602.
- Pennekamp, P., Batstone, R. J., & Mansour, J. K. (2019, December). *Eyewitness identification confidence: Requesting, articulating, and apperceiving*. Poster at the Scottish Institute for Policing Research.
- Proffitt, D. (2006). Embodied perception and the economy of action. *Perspectives on Psychological Science*, 1(2), 110-122.

- 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.
- Rhead, L. M., Rodriguez, D. N., Korobeynikov, V., Yip, J. H., & Bull Kovera, M. (2015). The effects of lineup administrator influence and mortality salience on witness identification accuracy. *Journal of Forensic Psychology Practice, 15*(3), 248-274.
- Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J. C., & Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC bioinformatics, 12*(1), 77.
- Robinson, M. D., & Johnson, J. T. (1996). Recall Memory, Recognition Memory, and the Eyewitness Confidence-Accuracy Correlation. *Journal of Applied Psychology, 81*(5), 587–594. <https://doi.org/10.1037/0021-9010.81.5.587>
- Rotello, C. M. (2000). Recall processes in recognition memory. In *Psychology of Learning and Motivation* (Vol. 40, pp. 183-221). Academic Press.
- Sauer, J. D., Brewer, N., & Weber, N. (2008). Multiple confidence estimates as indices of eyewitness memory. *Journal of Experimental Psychology: General, 137*(3), 528.
- Saur, Brewer, Zweck, & Weber. (2010). The effect of retention interval on the confidence-accuracy relationship for eyewitness identification. *Law and Human Behavior, 34*, 337-347. DOI 10.1007/s10979-009-9192-x.
- Semmler, C., Dunn, J., Mickes, L., & Wixted, J. T. (2018). The role of estimator variables in eyewitness identification. *Journal of Experimental Psychology: Applied, 24*(3), 400.

- Shah, Gaule, Sowden, Bird, & Cook. (2015). The 20-item prosopagnosia index (PI20): A self-report instrument for identifying developmental prosopagnosia. *Royal Society Open Science*, 2(6), 140343. <https://doi.org/10.1098/rsos.140343/>.
- Smith, A., Wells, G., Lindsay, R.C.L., & Penrod, S. (2016). Fair lineups are better than biased lineups and showups, but not because they increase underlying discriminability. *Law and Human Behavior*, 41(2), 127.
- Stantić, M., Pounder, Z., Bate, S., Susilo, T., Catmur, C., & Bird, G. (2022). Individuals with developmental prosopagnosia show independent impairments in face perception, face memory and face matching. *cortex*, 157, 266-273.
- Sterne, J. A. C., Savović, J., Page, M. J., Elbers, R. G., Blencowe, N. S., Boutron, I., Cates, C. J., Cheng, H. Y., Corbett, M. S., Eldridge, S. M., Emberson, J. R., Hernán, M. A., Hopewell, S., Hróbjartsson, A., Junqueira, D. R., Jüni, P., Kirkham, J. J., Lasserson, T., Li, T., McAleenan, A., ... Higgins, J. P. T. (2019). RoB 2: a revised tool for assessing risk of bias in randomised trials. *BMJ (Clinical research ed.)*, 366, 14898. <https://doi.org/10.1136/bmj.14898>
- Strüber, D., & Stadler, M. (1999). Differences in top—down influences on the reversal rate of different categories of reversible figures. *Perception*, 28(10), 1185-1196.
- Sučić, I., Tokić, D., & Ivešić, M. (2015). Field study of response accuracy and decision confidence with regard to lineup composition and lineup presentation. *Psychology, Crime & Law*, 21(8), 798-819.
- Tanaka, & Farah. (1993). Parts and wholes in face recognition. *The Quarterly Journal of Experimental Psychology Section A*, 46(2), 225-245.

- Tredoux, C. G. (1998). Statistical inference on measures of lineup fairness. *Law and Human Behavior*, 22, 217–237. doi:10.1023/A:1025746220886.
- Tredoux, C. (1999). Statistical considerations when determining measures of lineup size and lineup bias. *Applied Cognitive Psychology*, 13, S9–S26. doi: 10.1002/(SICI)1099-0720(199911)13:1+3.0.CO;2-1
- Tunnicliff, J. L., & Clark, S. E. (2000). Selecting foils for identification lineups: Matching suspects or descriptions?. *Law and Human Behavior*, 24(2), 231-258.
- Turtle, J. W., Lindsay, R. C. L., & Wells, G. L. (2003). Best practice recommendations for eyewitness evidence procedures: New ideas for the oldest way to solve a case. *The Canadian Journal of Police and Security Services*, 1(1), 5-18.
- Vieira, Bottino, Laurentini, De Simone. (2014). Detecting Siblings in Image Pairs, *The Visual Computer*, 30(12), p. 1333-1345, doi: 10.1007/s00371-013-0884-3
- Walker-Smith, G. (1978). The effects of delay and exposure duration in a face recognition task. *Perception & Psychophysics*, 24(1), 63-70.
- 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.
- Wells, G. L., Leippe, M. R., & Ostrom, T. M. (1979). Guidelines for empirically assessing the fairness of a lineup. *Law and Human Behavior*, 3(4), 285-293.
- Wells & Olson. (2003). Eyewitness testimony. *Annual review of Psychology*, 54(1), 277-295.
- Wells, G. L., Rydell, S. M., & Seelau, E. P. (1993). The selection of distractors for eyewitness lineups. *Journal of Applied Psychology*, 78(5), 835.

- Wells, G. L., Small, M., Penrod, S., Malpass, R. S., Fulero, S. M., & Brimacombe, C. E. (1998). Eyewitness identification procedures: Recommendations for lineups and photospreads. *Law and Human Behavior*, 22(6), 603.
- Wixted, J. T., & Mickes, L. (2014). A signal-detection-based diagnostic-feature-detection model of eyewitness identification. *Psychological Review*, 121(2), 262.
- 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.
- Wixted, J. T., & Wells, G. L. (2017). The relationship between eyewitness confidence and identification accuracy: A new synthesis. *Psychological Science in the Public Interest*, 18(1), 10-65.

Appendices

Appendix A

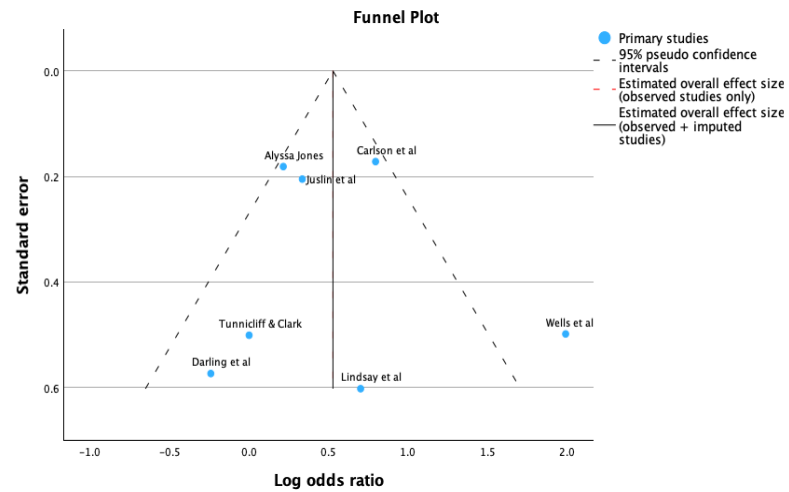
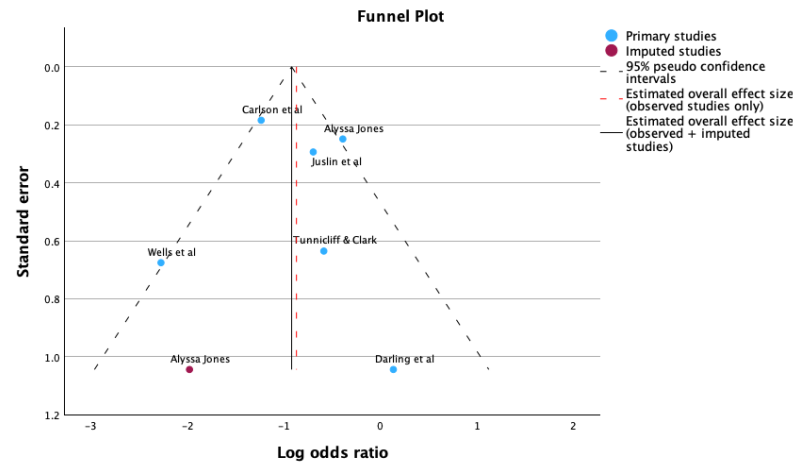
Figure 1*Funnel Plot for Correct IDs for Target Present Lineups***Figure 2***Funnel Plot for Filler IDs for Target Present Lineups*

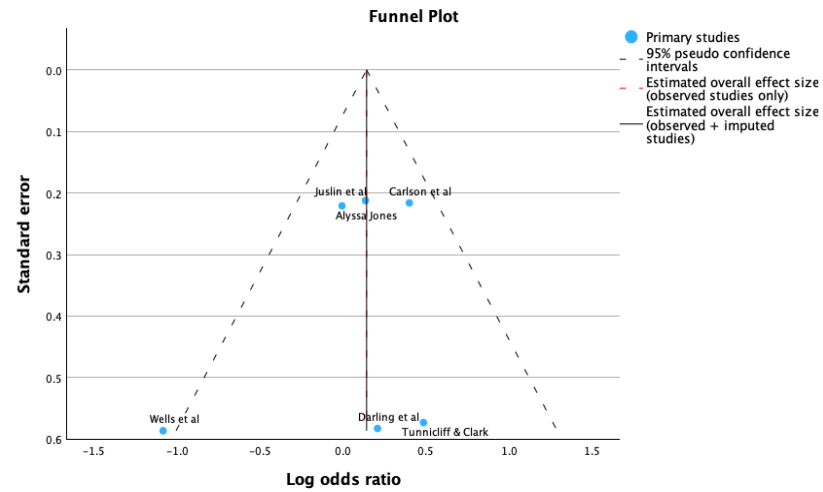
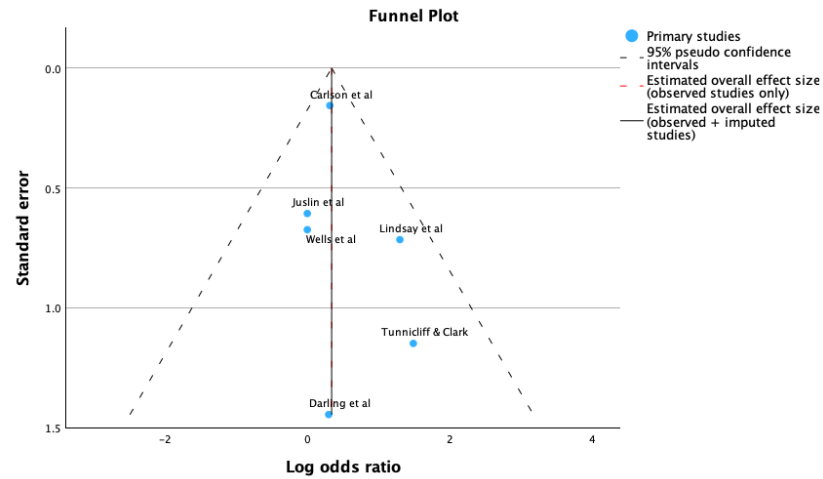
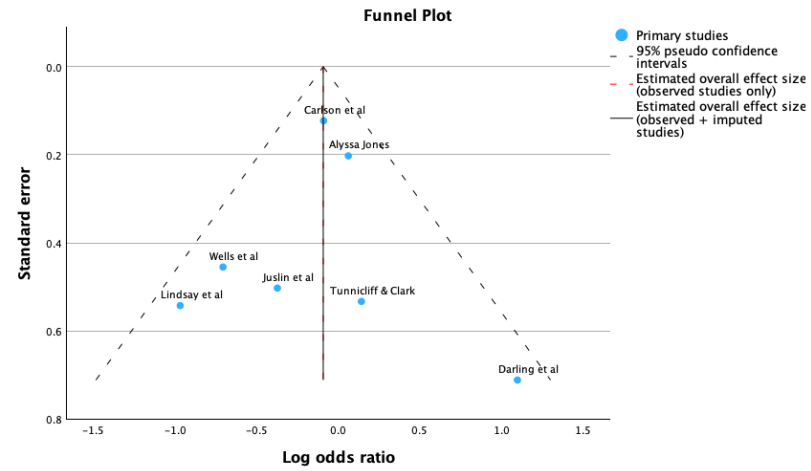
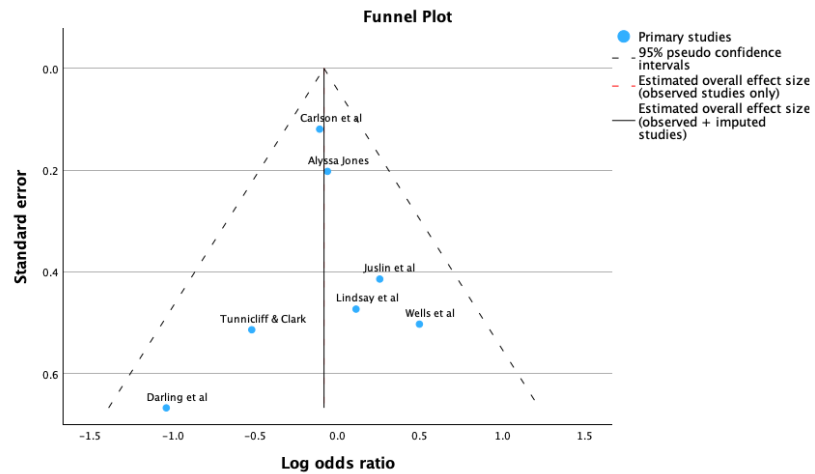
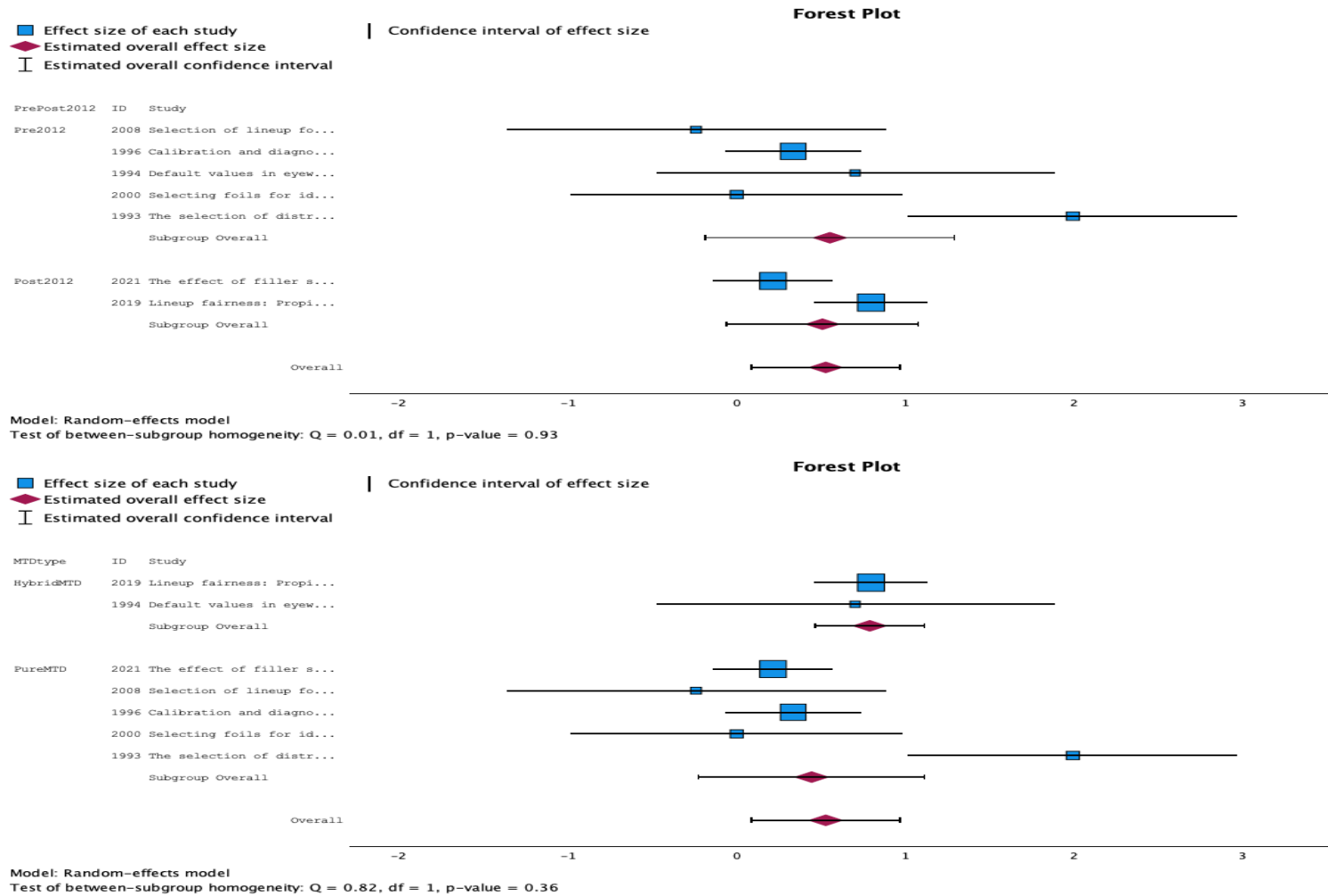
Figure 3*Funnel Plot for Non-IDs for Target Present Lineups***Figure 4***Funnel Plot for Innocent Suspect IDs for Target Absent Lineups*

Figure 5*Funnel Plot for Filler IDs for Target Absent Lineups***Figure 6***Funnel Plot for Non-IDs for Target Absent Lineups*

Appendix B

Figure 1

Forest Plot Depicting Moderator Analyses for Correct IDs for Target Present Lineups



■ Effect size of each study
◆ Estimated overall effect size
| Estimated overall confidence interval

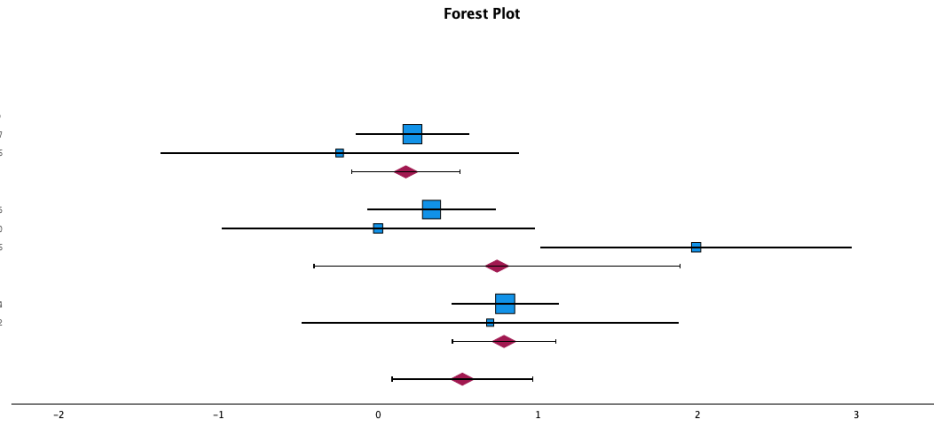
SubsequentFilter ID	Study LogOR Std. Error Lower Upper p-value Weight Weight (%)
Unclear	The effect of filler s... 2021 0.22 0.18 -0.14 0.57 0.24 4.02 20.27
	Selection of lineup fo... 2008 -0.24 0.57 -1.37 0.88 0.67 1.83 9.26
	Subgroup Overall 0.17 0.17 -0.17 0.51 0.31
No	Calibration and diagno... 1996 0.33 0.21 -0.07 0.74 0.10 3.87 19.55
	Selecting foils for id... 2000 -0.00 0.50 -0.98 0.98 1.00 2.14 10.80
	The selection of distr... 1993 1.99 0.50 1.02 2.97 0.00 2.15 10.86
	Subgroup Overall 0.74 0.58 -0.40 1.89 0.20
Yes	Lineup fairness: Propi... 2019 0.80 0.17 0.46 1.13 0.00 4.07 20.54
	Default values in eyew... 1994 0.70 0.60 -0.48 1.88 0.24 1.73 8.72
	Subgroup Overall 0.79 0.17 0.47 1.11 0.00
Overall	0.53 0.22 0.09 0.97 0.02

Model: Random-effects model

Heterogeneity: Tau-squared = 0.22, H-squared = 3.72, I-squared = 0.73

Homogeneity: Q = 17.82, df = 6, p-value = 0.01

Test of between-subgroup homogeneity: Q = 6.79, df = 2, p-value = 0.03



■ Effect size of each study
◆ Estimated overall effect size
| Estimated overall confidence interval

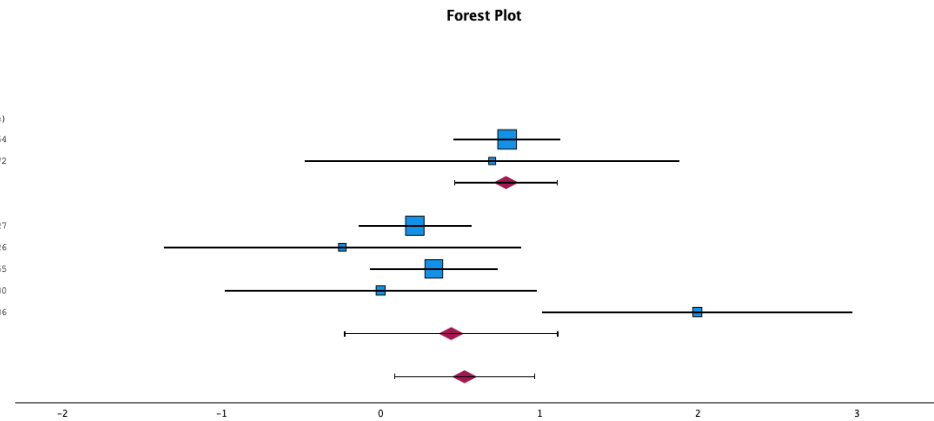
ParticipantType ID	Study LogOR Std. Error Lower Upper p-value Weight Weight (%)
Community Volunteer	Lineup fairness: Propi... 2019 0.80 0.17 0.46 1.13 0.00 4.07 20.54
	Default values in eyew... 1994 0.70 0.60 -0.48 1.88 0.24 1.73 8.72
	Subgroup Overall 0.79 0.17 0.47 1.11 0.00
University Student	The effect of filler s... 2021 0.22 0.18 -0.14 0.57 0.24 4.02 20.27
	Selection of lineup fo... 2008 -0.24 0.57 -1.37 0.88 0.67 1.83 9.26
	Calibration and diagno... 1996 0.33 0.21 -0.07 0.74 0.10 3.87 19.55
	Selecting foils for id... 2000 -0.00 0.50 -0.98 0.98 1.00 2.14 10.80
	The selection of distr... 1993 1.99 0.50 1.02 2.97 0.00 2.15 10.86
	Subgroup Overall 0.44 0.34 -0.23 1.11 0.19
Overall	0.53 0.22 0.09 0.97 0.02

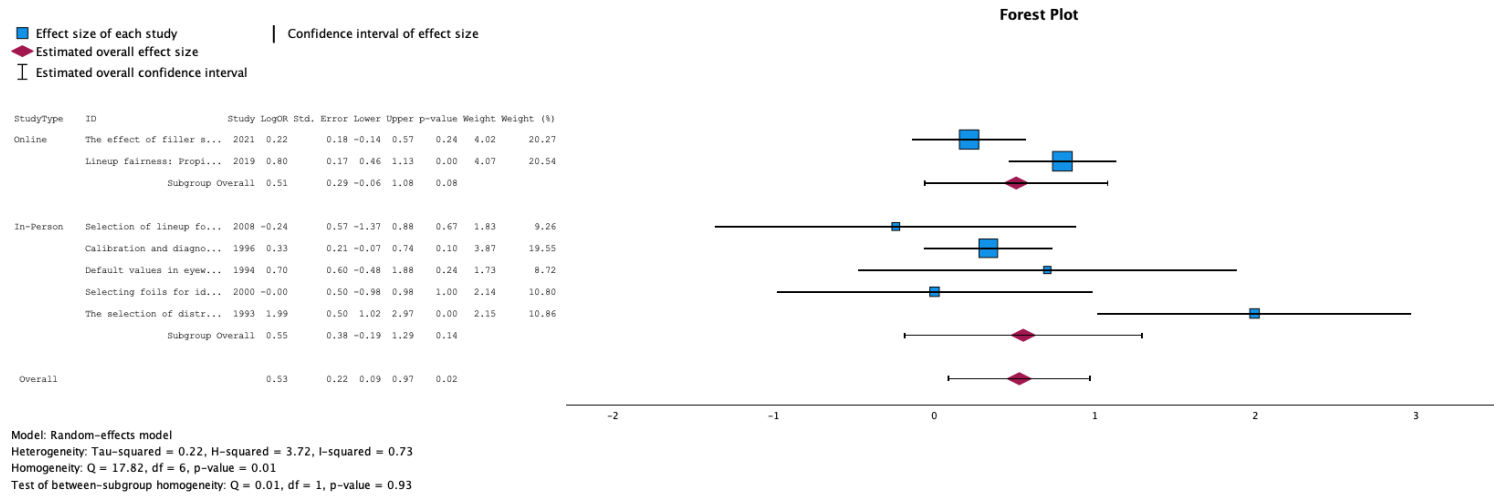
Model: Random-effects model

Heterogeneity: Tau-squared = 0.22, H-squared = 3.72, I-squared = 0.73

Homogeneity: Q = 17.82, df = 6, p-value = 0.01

Test of between-subgroup homogeneity: Q = 0.82, df = 1, p-value = 0.36

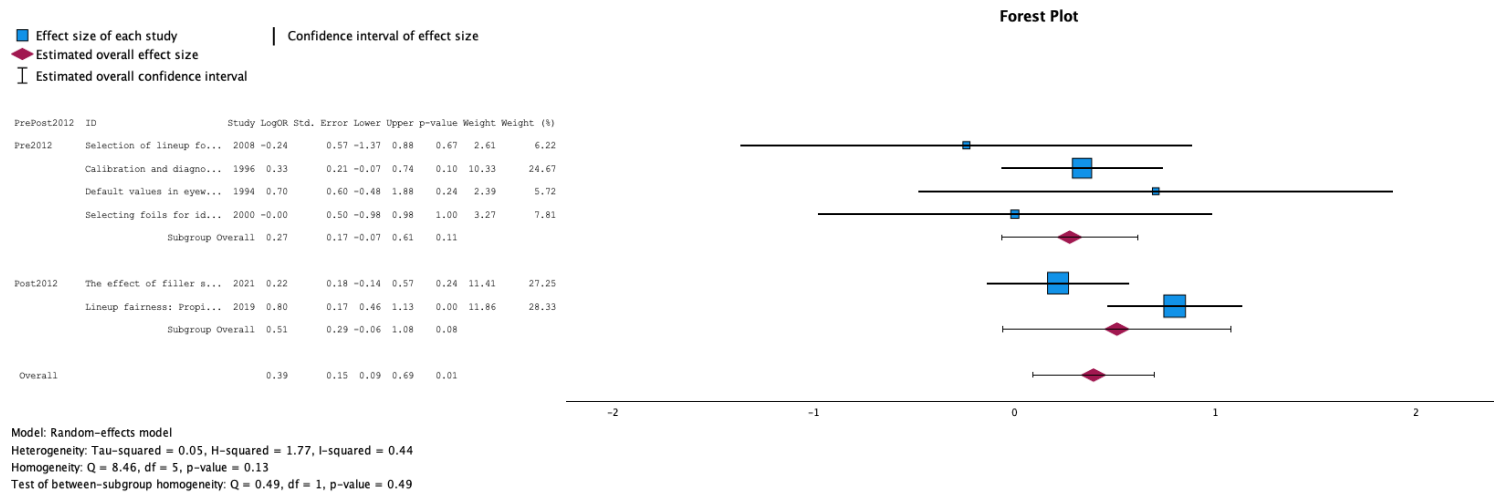




Note: Wells (1993) outlier is included in forest plot.

Figure 2

Forest Plot Depicting Moderator Analyses for Correct IDs for Target Present Lineups



■ Effect size of each study | Confidence interval of effect size
◆ Estimated overall effect size
| Estimated overall confidence interval

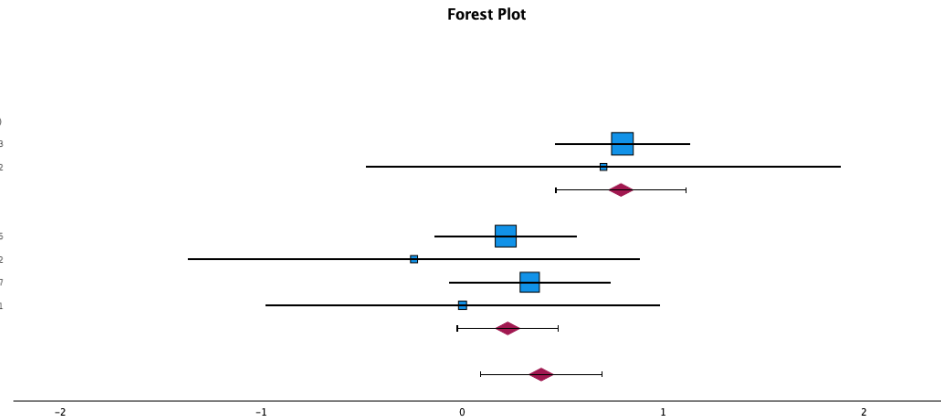
MTDtype	ID	Study	LogOR	Std. Error	Lower	Upper	p-value	Weight	Weight (%)
HybridMTD	Lineup fairness: Propi...	2019	0.80	0.17	0.46	1.13	0.00	11.86	28.33
	Default values in eyew...	1994	0.70	0.60	-0.48	1.88	0.24	2.39	5.72
	Subgroup Overall	0.79	0.17	0.47	1.11	0.00			
PureMTD	The effect of filler s...	2021	0.22	0.18	-0.14	0.57	0.24	11.41	27.25
	Selection of lineup fo...	2008	-0.24	0.57	-1.37	0.88	0.67	2.61	6.22
	Calibration and diagno...	1996	0.33	0.21	-0.07	0.74	0.10	10.33	24.67
	Selecting foils for id...	2000	-0.00	0.50	-0.98	0.98	1.00	3.27	7.81
	Subgroup Overall	0.22	0.13	-0.03	0.48	0.08			
Overall		0.39	0.15	0.09	0.69	0.01			

Model: Random-effects model

Heterogeneity: Tau-squared = 0.05, H-squared = 1.77, I-squared = 0.44

Homogeneity: Q = 8.46, df = 5, p-value = 0.13

Test of between-subgroup homogeneity: Q = 7.29, df = 1, p-value = 0.01



■ Effect size of each study | Confidence interval of effect size
◆ Estimated overall effect size
| Estimated overall confidence interval

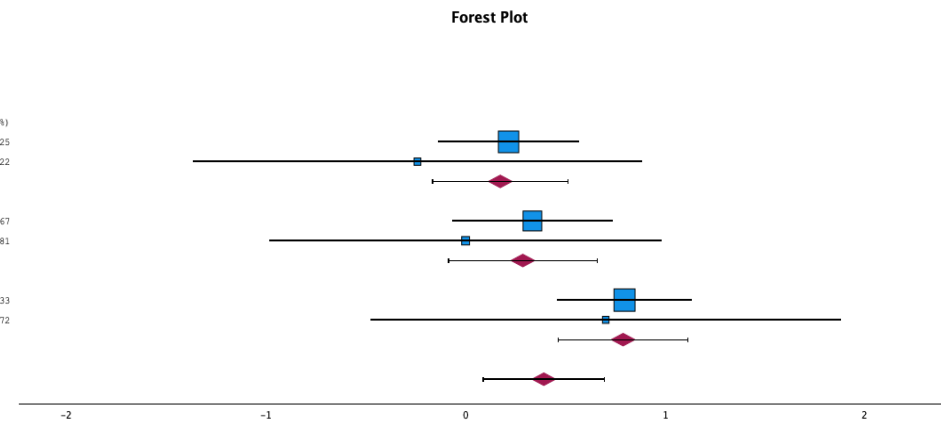
SubsequentFilter	ID	Study	LogOR	Std. Error	Lower	Upper	p-value	Weight	Weight (%)
Unclear	The effect of filler s...	2021	0.22	0.18	-0.14	0.57	0.24	11.41	27.25
	Selection of lineup fo...	2008	-0.24	0.57	-1.37	0.88	0.67	2.61	6.22
	Subgroup Overall	0.17	0.17	-0.17	0.51	0.31			
No	Calibration and diagno...	1996	0.33	0.21	-0.07	0.74	0.10	10.33	24.67
	Selecting foils for id...	2000	-0.00	0.50	-0.98	0.98	1.00	3.27	7.81
	Subgroup Overall	0.29	0.19	-0.09	0.66	0.13			
Yes	Lineup fairness: Propi...	2019	0.80	0.17	0.46	1.13	0.00	11.86	28.33
	Default values in eyew...	1994	0.70	0.60	-0.48	1.88	0.24	2.39	5.72
	Subgroup Overall	0.79	0.17	0.47	1.11	0.00			
Overall		0.39	0.15	0.09	0.69	0.01			

Model: Random-effects model

Heterogeneity: Tau-squared = 0.05, H-squared = 1.77, I-squared = 0.44

Homogeneity: Q = 8.46, df = 5, p-value = 0.13

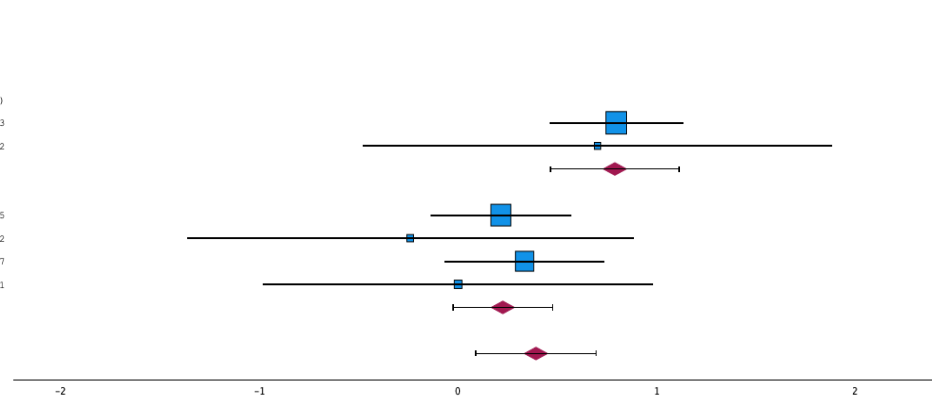
Test of between-subgroup homogeneity: Q = 7.48, df = 2, p-value = 0.02



ParticipantType	ID	Study	LogOR	Std. Error	Lower	Upper	p-value	Weight	Weight (%)
Community Volunteer	Lineup fairness: Propi...	2019	0.80	0.17	0.46	1.13	0.00	11.86	28.33
	Default values in eyew...	1994	0.70	0.60	-0.48	1.88	0.24	2.39	5.72
	Subgroup Overall	0.79	0.17	0.47	1.11	0.00			
University Student	The effect of filler s...	2021	0.22	0.18	-0.14	0.57	0.24	11.41	27.25
	Selection of lineup fo...	2008	-0.24	0.57	-1.37	0.88	0.67	2.61	6.22
	Calibration and diagno...	1996	0.33	0.21	-0.07	0.74	0.10	10.33	24.67
	Selecting foils for id...	2000	-0.00	0.50	-0.98	0.98	1.00	3.27	7.81
	Subgroup Overall	0.22	0.13	-0.03	0.48	0.08			
Overall		0.39	0.15	0.09	0.69	0.01			

Model: Random-effects model
Heterogeneity: Tau-squared = 0.05, H-squared = 1.77, I-squared = 0.44
Homogeneity: Q = 8.46, df = 5, p-value = 0.13
Test of between-subgroup homogeneity: Q = 7.29, df = 1, p-value = 0.01

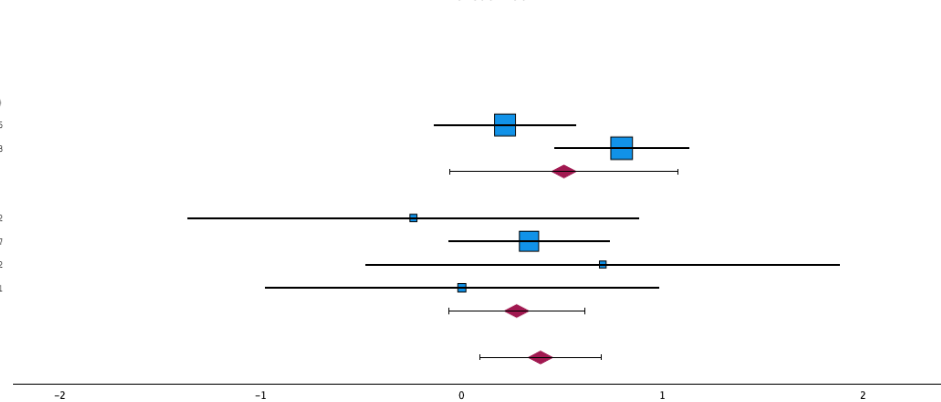
Forest Plot



StudyType	ID	Study	LogOR	Std. Error	Lower	Upper	p-value	Weight	Weight (%)
Online	The effect of filler s...	2021	0.22	0.18	-0.14	0.57	0.24	11.41	27.25
	Lineup fairness: Propi...	2019	0.80	0.17	0.46	1.13	0.00	11.86	28.33
	Subgroup Overall	0.51	0.29	-0.06	1.08	0.08			
In-Person	Selection of lineup fo...	2008	-0.24	0.57	-1.37	0.88	0.67	2.61	6.22
	Calibration and diagno...	1996	0.33	0.21	-0.07	0.74	0.10	10.33	24.67
	Default values in eyew...	1994	0.70	0.60	-0.48	1.88	0.24	2.39	5.72
	Selecting foils for id...	2000	-0.00	0.50	-0.98	0.98	1.00	3.27	7.81
	Subgroup Overall	0.27	0.17	-0.07	0.61	0.11			
Overall		0.39	0.15	0.09	0.69	0.01			

Model: Random-effects model
Heterogeneity: Tau-squared = 0.05, H-squared = 1.77, I-squared = 0.44
Homogeneity: Q = 8.46, df = 5, p-value = 0.13
Test of between-subgroup homogeneity: Q = 0.49, df = 1, p-value = 0.49

Forest Plot

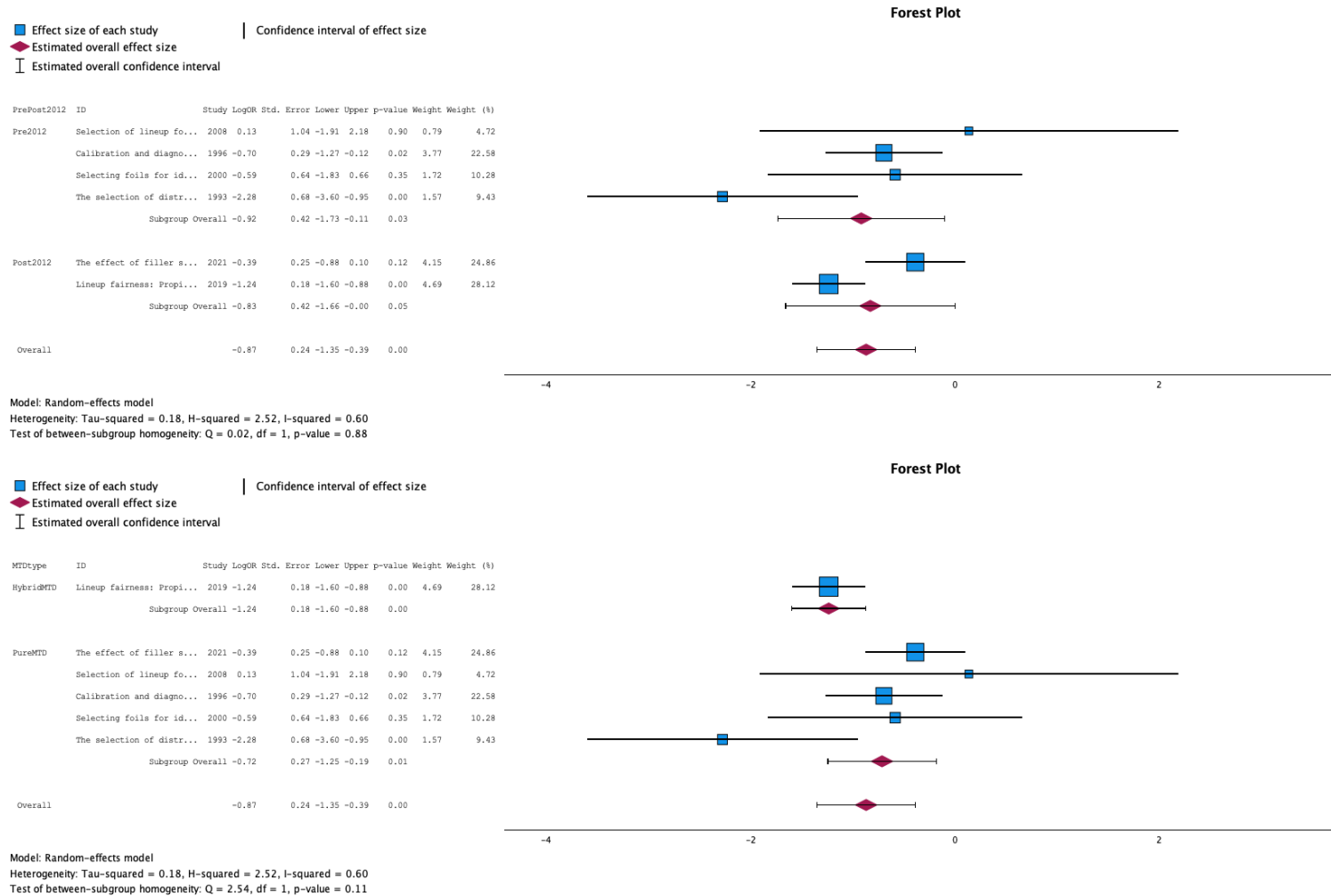


Note: Forest plot displaying the moderator analysis results with Wells (1993) outlier removed

Appendix C

Figure 1

Forest Plot Depicting Moderator Analyses for Filler IDs for Target Present Lineups



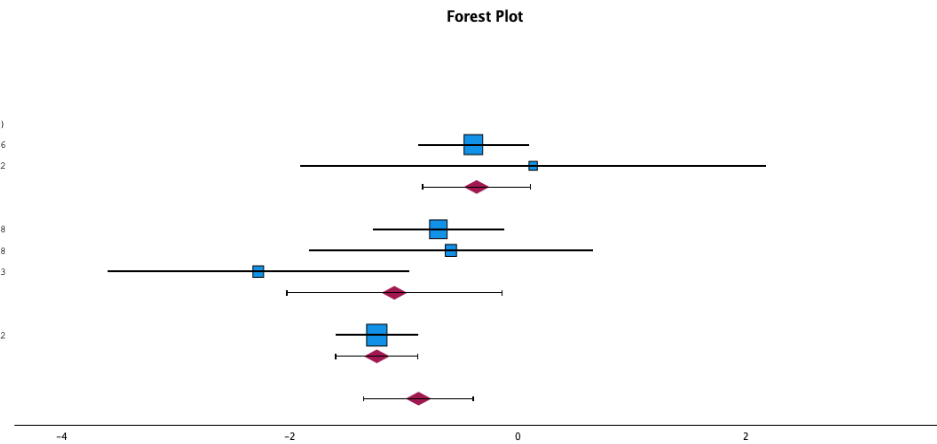
■ Effect size of each study | Confidence interval of effect size
◆ Estimated overall effect size
| Estimated overall confidence interval

SubsequentFilter ID	Study LogOR Std. Error Lower Upper p-value Weight Weight (%)
Unclear	The effect of filler s... 2021 -0.39 0.25 -0.88 0.10 0.12 4.15 24.86
	Selection of lineup fo... 2008 0.13 1.04 -1.91 2.18 0.90 0.79 4.72
	Subgroup Overall -0.36 0.24 -0.84 0.11 0.13
No	Calibration and diagno... 1996 -0.70 0.29 -1.27 -0.12 0.02 3.77 22.58
	Selecting foils for id... 2000 -0.59 0.64 -1.83 0.66 0.35 1.72 10.28
	The selection of distr... 1993 -2.28 0.68 -3.60 -0.95 0.00 1.57 9.43
	Subgroup Overall -1.08 0.48 -2.03 -0.14 0.02
Yes	Lineup fairness: Propi... 2019 -1.24 0.18 -1.60 -0.88 0.00 4.69 28.12
	Subgroup Overall -1.24 0.18 -1.60 -0.88 0.00
Overall	-0.87 0.24 -1.35 -0.39 0.00

Model: Random-effects model

Heterogeneity: Tau-squared = 0.18, H-squared = 2.52, I-squared = 0.60

Test of between-subgroup homogeneity: Q = 8.38, df = 2, p-value = 0.02



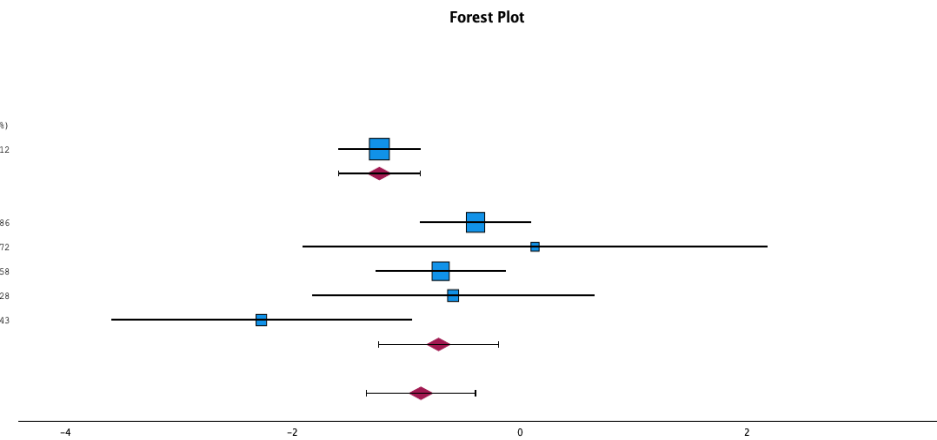
■ Effect size of each study | Confidence interval of effect size
◆ Estimated overall effect size
| Estimated overall confidence interval

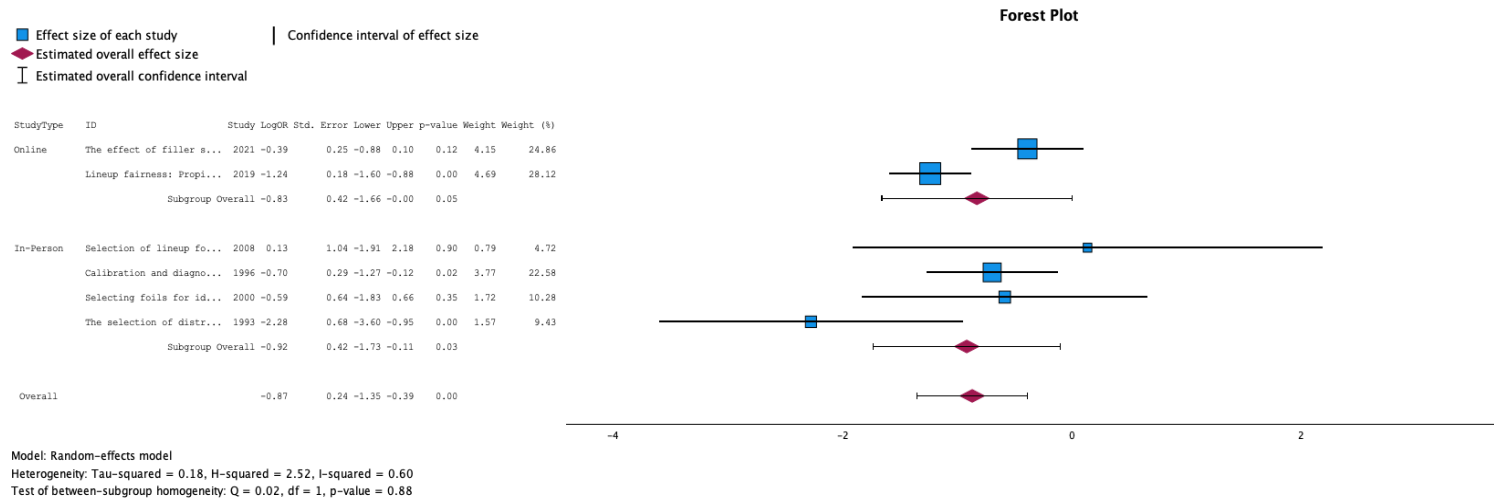
ParticipantType ID	Study LogOR Std. Error Lower Upper p-value Weight Weight (%)
Community Volunteer	Lineup fairness: Propi... 2019 -1.24 0.18 -1.60 -0.88 0.00 4.69 28.12
	Subgroup Overall -1.24 0.18 -1.60 -0.88 0.00
University Student	The effect of filler s... 2021 -0.39 0.25 -0.88 0.10 0.12 4.15 24.86
	Selection of lineup fo... 2008 0.13 1.04 -1.91 2.18 0.90 0.79 4.72
	Calibration and diagno... 1996 -0.70 0.29 -1.27 -0.12 0.02 3.77 22.58
	Selecting foils for id... 2000 -0.59 0.64 -1.83 0.66 0.35 1.72 10.28
	The selection of distr... 1993 -2.28 0.68 -3.60 -0.95 0.00 1.57 9.43
	Subgroup Overall -0.72 0.27 -1.25 -0.19 0.01
Overall	-0.87 0.24 -1.35 -0.39 0.00

Model: Random-effects model

Heterogeneity: Tau-squared = 0.18, H-squared = 2.52, I-squared = 0.60

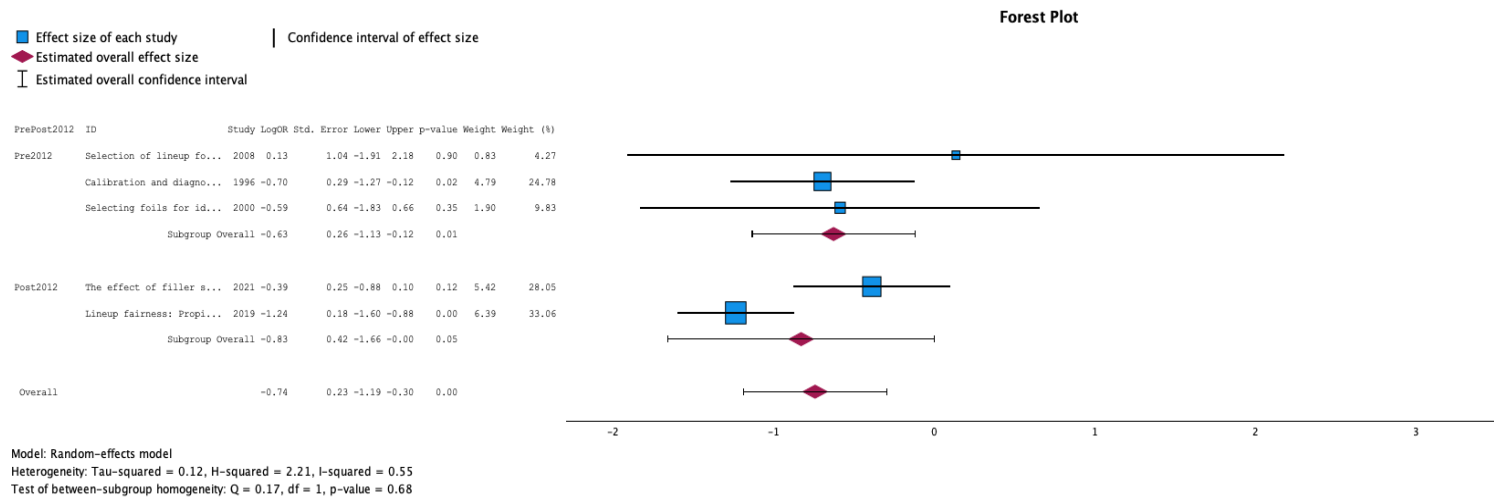
Test of between-subgroup homogeneity: Q = 2.54, df = 1, p-value = 0.11





Note: Wells (1993) outlier is included in forest plot.

Figure 2
Forest Plot Depicting Moderator Analyses for Filler IDs for Target Present Lineups

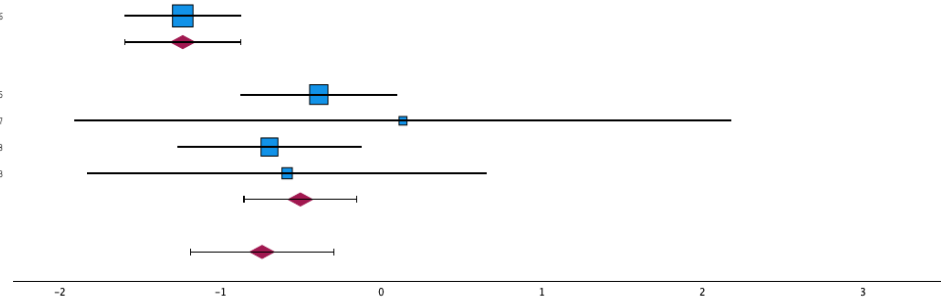


 Effect size of each study
 Estimated overall effect size
 Estimated overall confidence interval

Confidence interval of effect size

Forest Plot

MTDtype	ID	Study	LogOR	Std. Error	Lower	Upper	p-value	Weight	Weight (%)	
HybridMTD	Lineup fairness: Propi...	2019	-1.24	0.18	-1.60	-0.88	0.00	6.39	33.06	
		Subgroup Overall	-1.24	0.18	-1.60	-0.88	0.00			
PureMTD	The effect of filler s...	2021	-0.39	0.25	-0.88	0.10	0.12	5.42	28.05	
		Selection of lineup fo...	2008	0.13	1.04	-1.91	2.18	0.90	0.83	4.27
		Calibration and diagno...	1996	-0.70	0.29	-1.27	-0.12	0.02	4.79	24.78
		Selecting foils for id...	2000	-0.59	0.64	-1.83	0.66	0.35	1.90	9.83
		Subgroup Overall	-0.51	0.18	-0.86	-0.15	0.00			
Overall			-0.74	0.23	-1.19	-0.30	0.00			



Model: Random-effects model

Heterogeneity: Tau-squared = 0.12, H-squared = 2.21, I-squared = 0.55

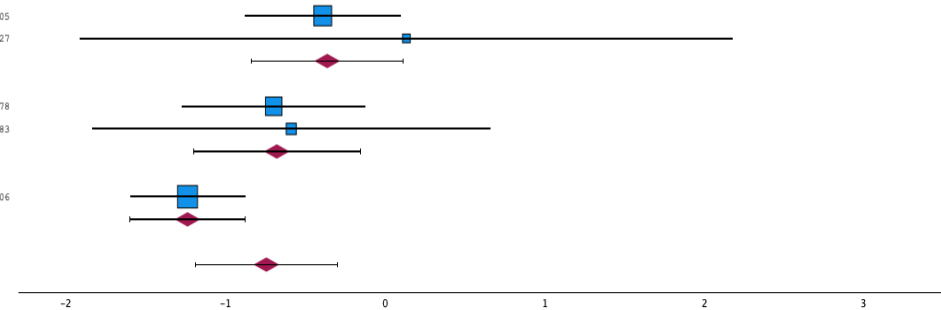
Test of between-subgroup homogeneity: Q = 8.12, df = 1, p-value = 0.00

 Effect size of each study
 Estimated overall effect size
 Estimated overall confidence interval

Confidence interval of effect size

Forest Plot

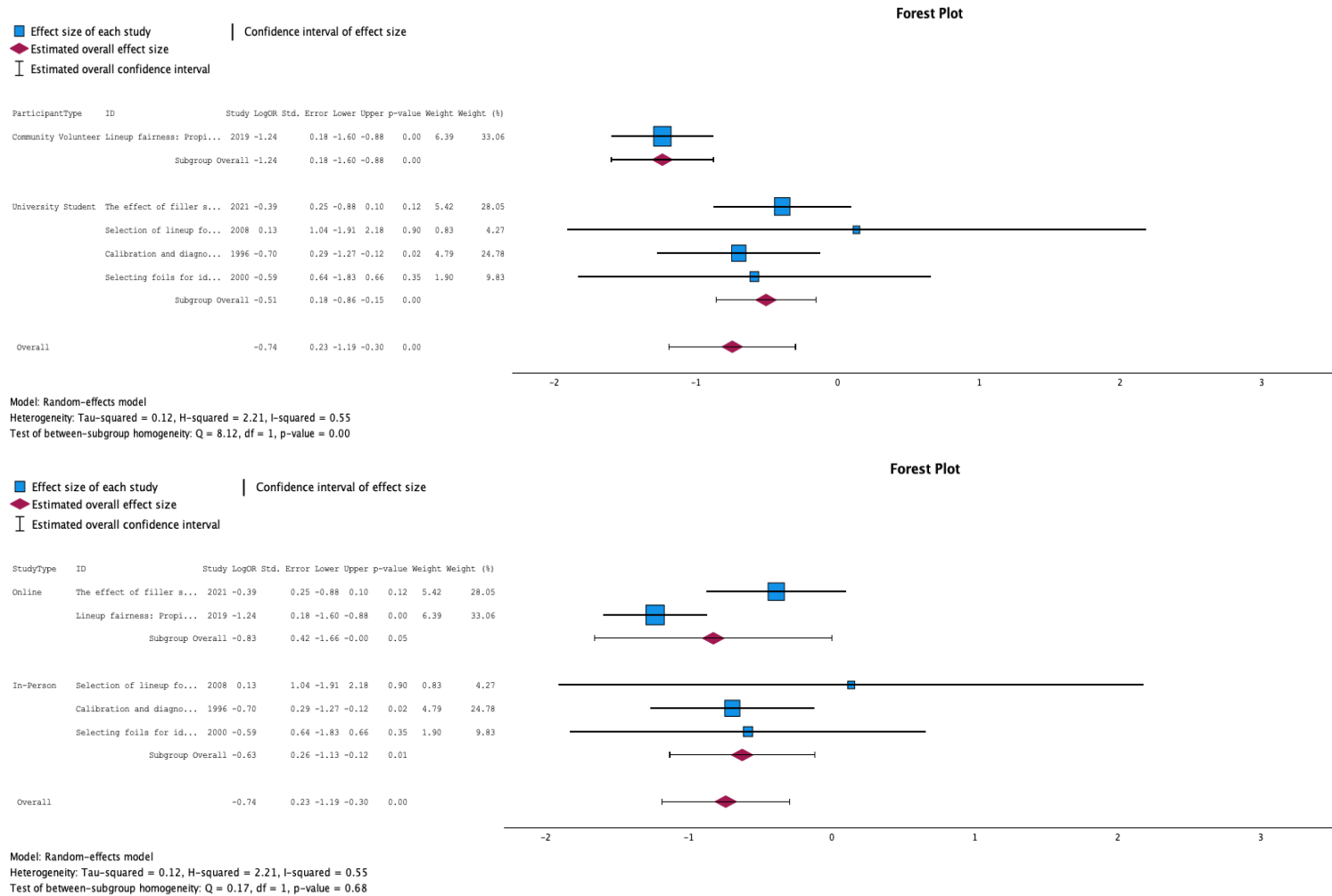
SubsequentFilter	ID	Study	LogOR	Std. Error	Lower	Upper	p-value	Weight	Weight (%)	
Unclear	The effect of filler s...	2021	-0.39	0.25	-0.88	0.10	0.12	5.42	28.05	
		Selection of lineup fo...	2008	0.13	1.04	-1.91	2.18	0.90	0.83	4.27
		Subgroup Overall	-0.36	0.24	-0.84	0.11	0.13			
No	Calibration and diagno...	1996	-0.70	0.29	-1.27	-0.12	0.02	4.79	24.78	
		Selecting foils for id...	2000	-0.59	0.64	-1.83	0.66	0.35	1.90	9.83
		Subgroup Overall	-0.68	0.27	-1.20	-0.16	0.01			
Yes	Lineup fairness: Propi...	2019	-1.24	0.18	-1.60	-0.88	0.00	6.39	33.06	
		Subgroup Overall	-1.24	0.18	-1.60	-0.88	0.00			
Overall			-0.74	0.23	-1.19	-0.30	0.00			



Model: Random-effects model

Heterogeneity: Tau-squared = 0.12, H-squared = 2.21, I-squared = 0.55

Test of between-subgroup homogeneity: Q = 8.88, df = 2, p-value = 0.01



Note: Forest plot displaying the moderator analysis results with Wells (1993) outlier removed.

Appendix D

Table 1

Test of Between-Subgroup Homogeneity for Non-IDs in Target Present Lineups

Moderators Examined	Findings
Pre-2012 vs post-2012	$Q = 0.31, df = 1, p = 0.58$
MTD type (Hybrid MTD vs pure MTD)	$Q = 2.01, df = 1, p = 0.16$
Subsequent Filtering (Unclear, yes, no)	$Q = 2.06, df = 2, p = 0.36$
Participant Type (Community volunteer vs university student)	$Q = 2.01, df = 1, p = 0.16$
Study Type (Online vs in-person)	$Q = 0.31, df = 1, p = 0.58$

Note: Wells (1993) outlier included in this table

Table 2

Test of Between-Subgroup Homogeneity for Non-IDs in Target Present Lineups

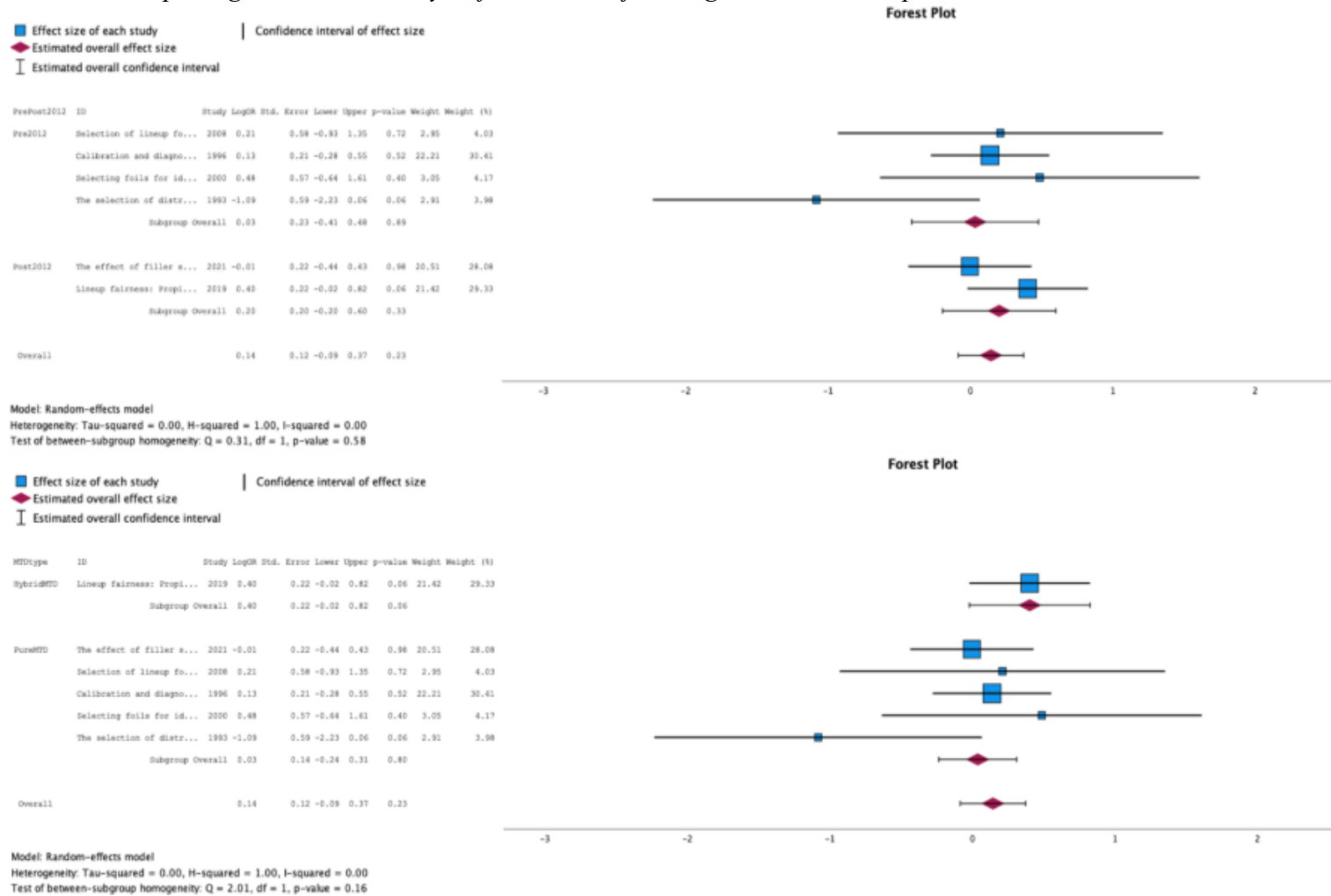
Moderators Examined	Findings
Pre-2012 vs post-2012	$Q = 0.00, df = 1, p = 0.95$
MTD type (Hybrid MTD vs pure MTD)	$Q = 1.32, df = 1, p = 0.25$
Subsequent Filtering (Unclear, yes, no)	$Q = 1.62, df = 2, p = 0.45$
Participant Type (Community volunteer vs university student)	$Q = 1.32, df = 1, p = 0.25$
Study Type (Online vs in-person)	$Q = 0.00, df = 1, p = 0.95$

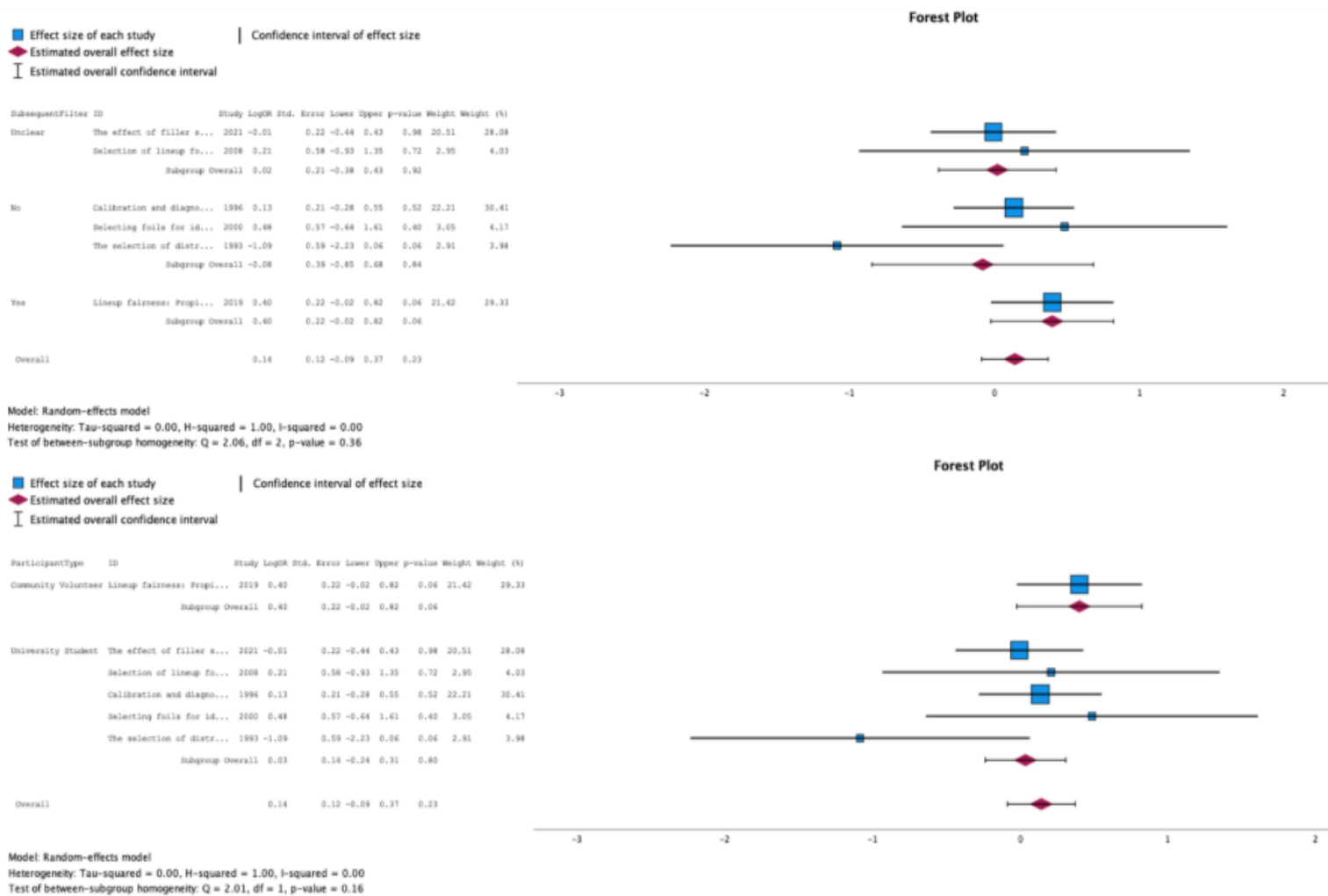
Note: Wells (1993) outlier **not** included in this table

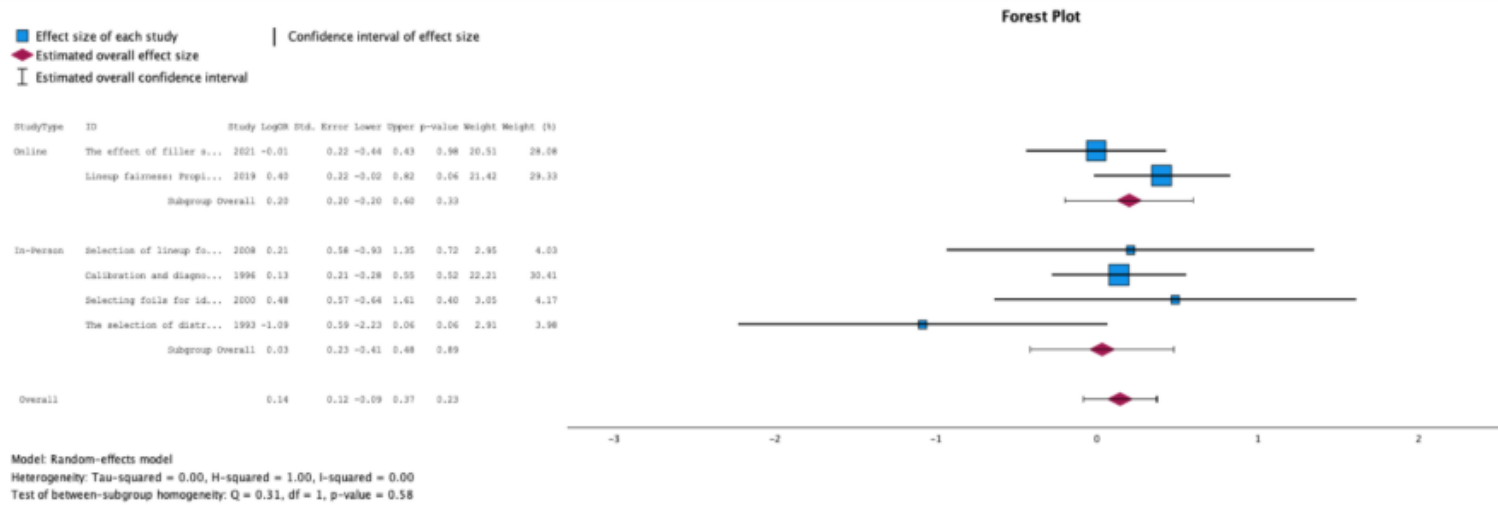
Appendix E

Figure 1

Forest Plot Depicting Moderator Analysis for Non-IDs for Target Present Lineups



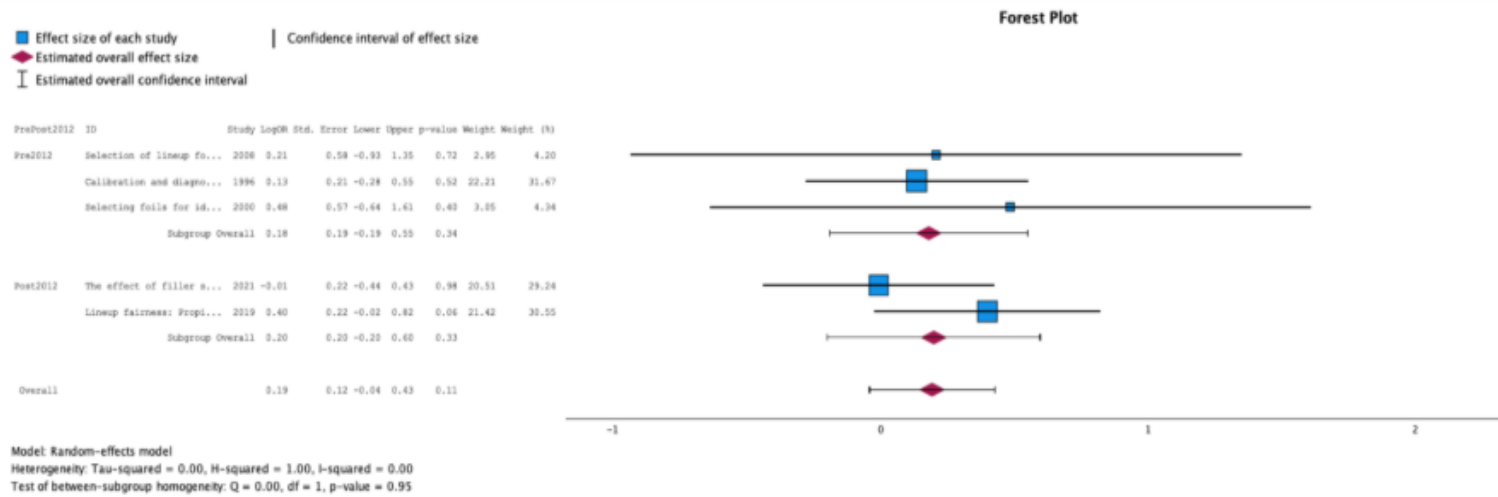


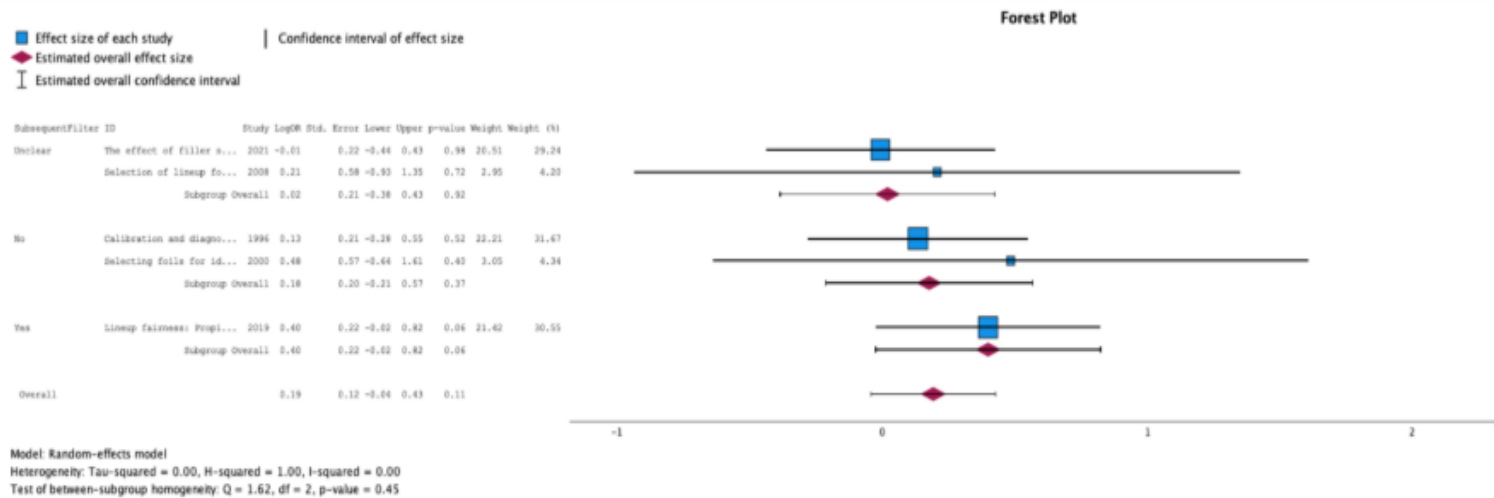
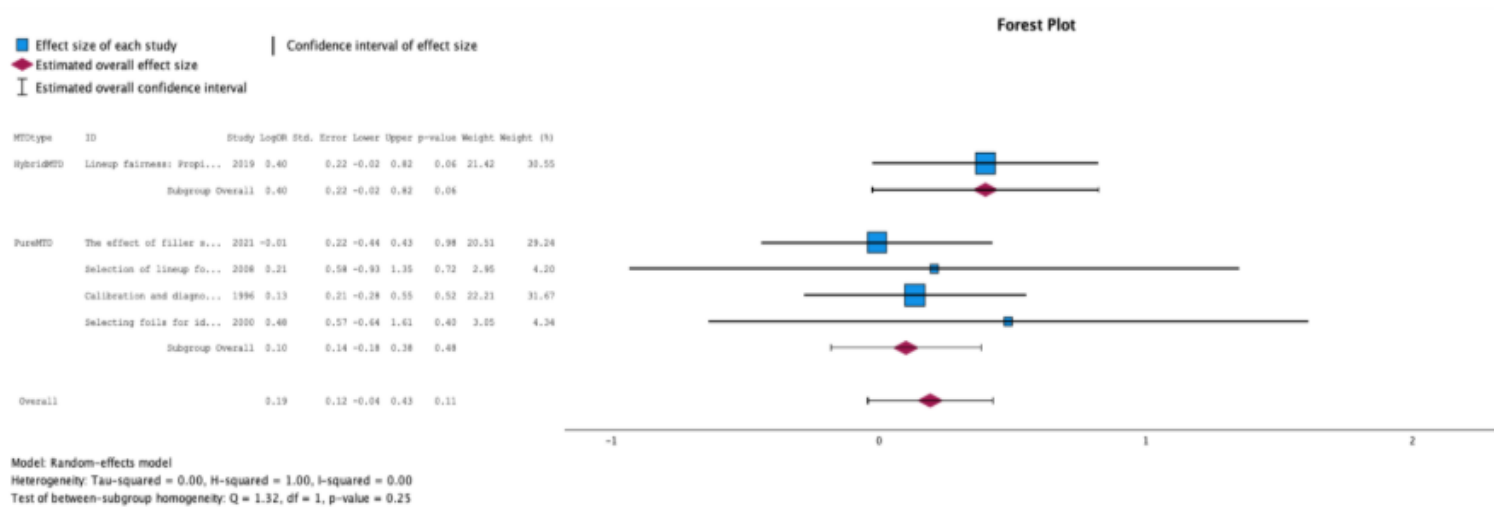


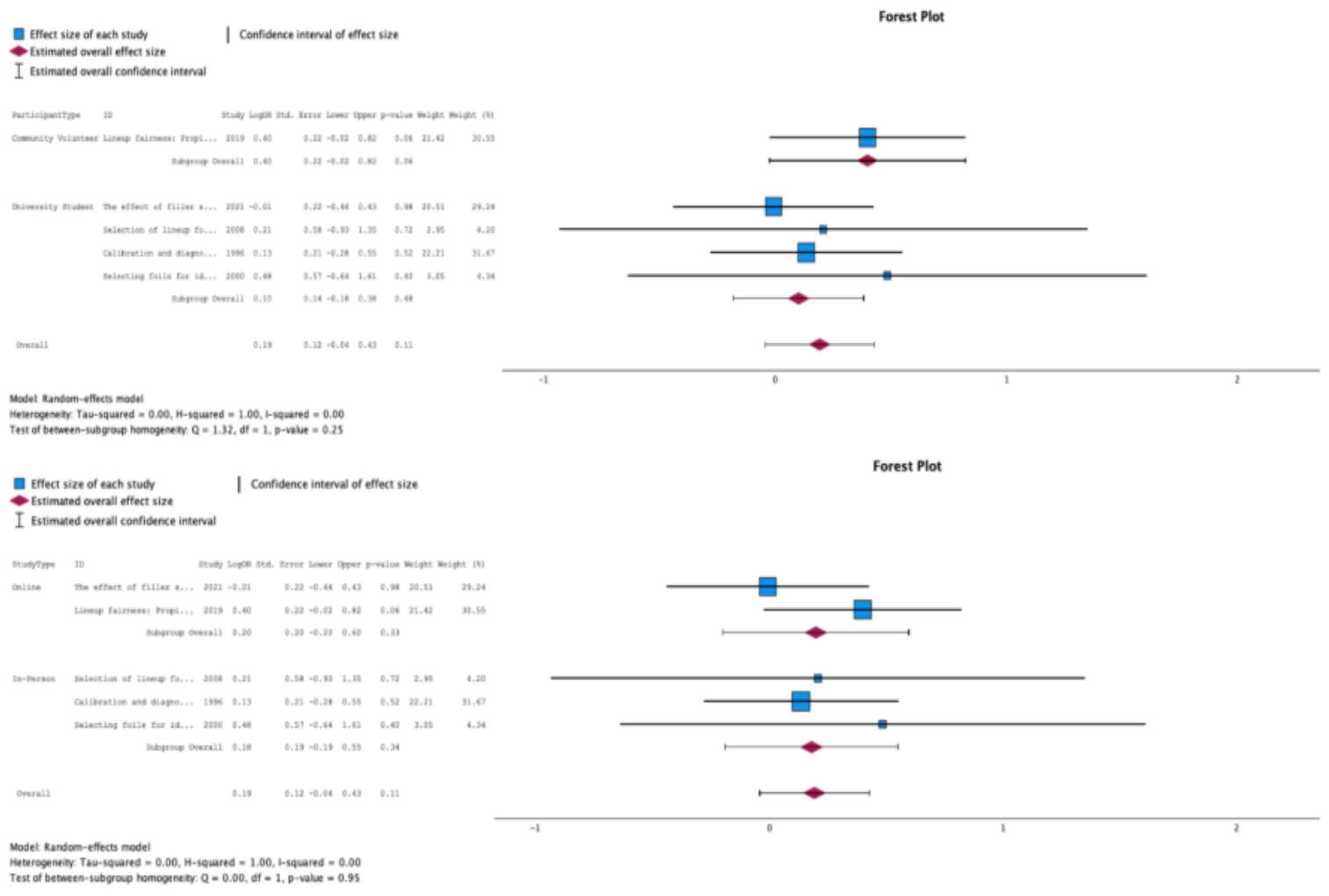
Note: Wells (1993) outlier is included in forest plot.

Figure 2

Forest Plot Depicting Moderator Analysis for Non-IDs for Target Present Lineups







Note: Forest plot displaying the moderator analysis results with Wells (1993) outlier removed.

Appendix F

Table 1

Test of Between-Subgroup Homogeneity for Innocent Suspect IDs in Target Absent Lineups

Moderators Examined	Findings
Pre-2012 vs post-2012	$Q = 0.16, df = 1, p = 0.69$
MTD type (Hybrid MTD vs pure MTD)	$Q = 0.36, df = 1, p = 0.55$
Subsequent Filtering (Unclear, yes, no)	$Q = 0.37, df = 2, p = 0.83$
Participant Type (Community volunteer vs university student)	$Q = 0.36, df = 1, p = 0.55$
Study Type (Online vs in-person)	$Q = 0.16, df = 1, p = 0.69$

Note: Wells (1993) outlier included in this table

Table 2

Test of Between-Subgroup Homogeneity for Innocent Suspect IDs in Target Absent Lineups

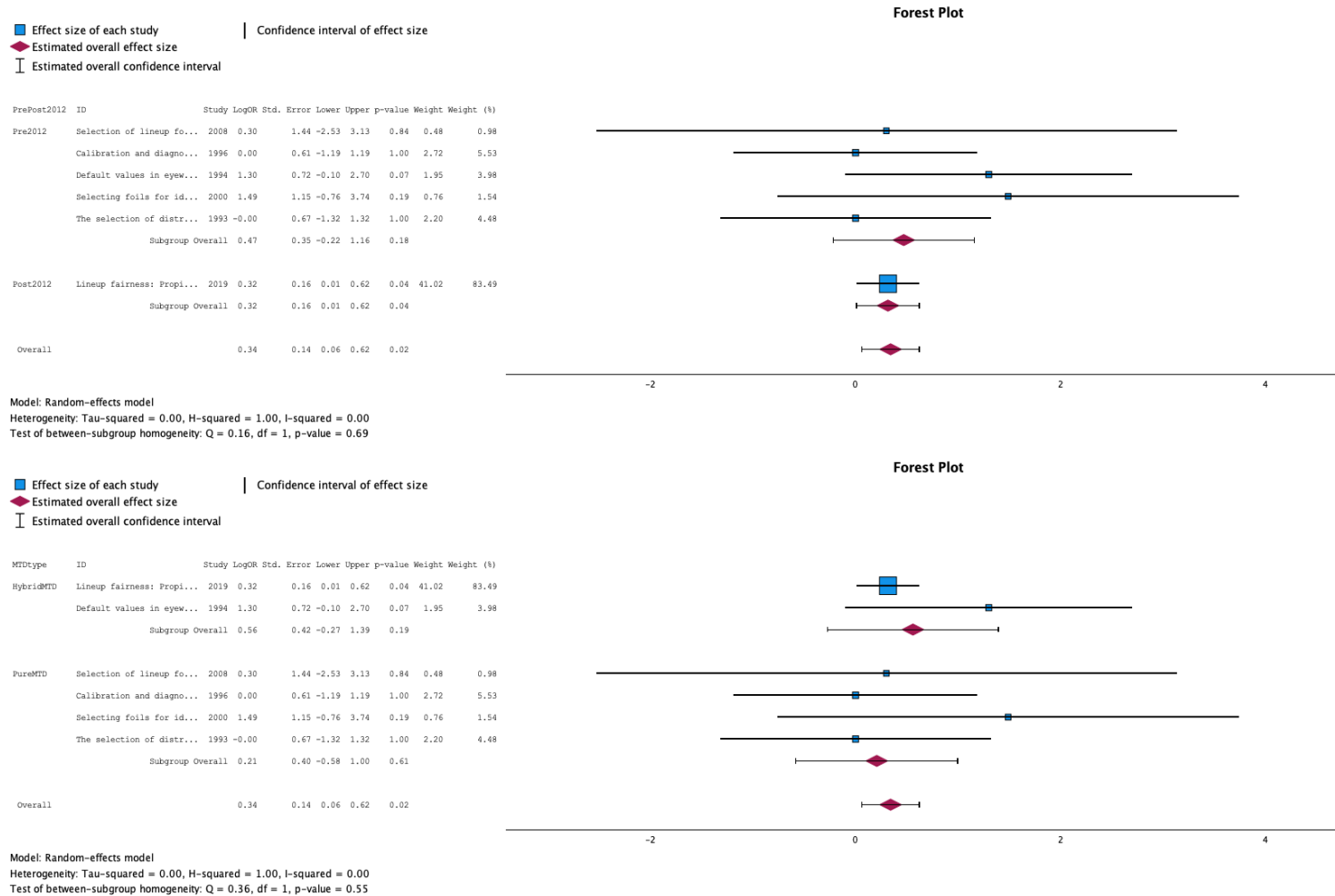
Moderators Examined	Findings
Pre-2012 vs post-2012	$Q = 0.55, df = 1, p = 0.46$
MTD type (Hybrid MTD vs pure MTD)	$Q = 0.13, df = 1, p = 0.72$
Subsequent Filtering (Unclear, yes, no)	$Q = 0.05, df = 2, p = 0.97$
Participant Type (Community volunteer vs university student)	$Q = 0.13, df = 1, p = 0.72$
Study Type (Online vs in-person)	$Q = 0.55, df = 1, p = 0.46$

Note: Wells (1993) outlier not included in this table

Appendix G

Figure 1

Forest Plot Depicting Moderator Analyses for Innocent Suspect IDs for Target Absent Lineups



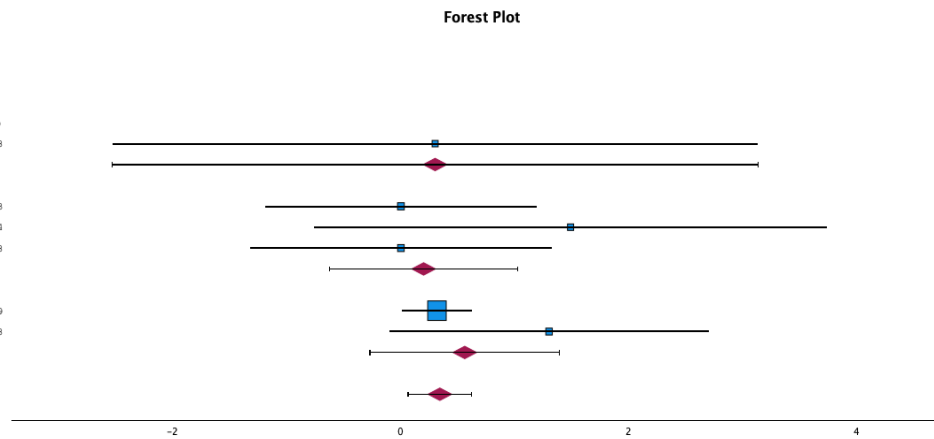
■ Effect size of each study
◆ Estimated overall effect size
| Estimated overall confidence interval

SubsequentFilter ID		Study LogOR	Std. Error	Lower	Upper	p-value	Weight	Weight (%)	
Unclear	Selection of lineup fo...	2008	0.30	1.44	-2.53	3.13	0.84	0.48	0.98
	Subgroup Overall	0.30		1.44	-2.53	3.13	0.84		
No	Calibration and diagno...	1996	0.00	0.61	-1.19	1.19	1.00	2.72	5.53
	Selecting foils for id...	2000	1.49	1.15	-0.76	3.74	0.19	0.76	1.54
	The selection of distr...	1993	-0.00	0.67	-1.32	1.32	1.00	2.20	4.48
	Subgroup Overall	0.20		0.42	-0.62	1.02	0.64		
Yes	Lineup fairness: Propi...	2019	0.32	0.16	0.01	0.62	0.04	41.02	83.49
	Default values in eyew...	1994	1.30	0.72	-0.10	2.70	0.07	1.95	3.98
	Subgroup Overall	0.56		0.42	-0.27	1.39	0.19		
Overall			0.34	0.14	0.06	0.62	0.02		

Model: Random-effects model

Heterogeneity: Tau-squared = 0.00, H-squared = 1.00, I-squared = 0.00

Test of between-subgroup homogeneity: $Q = 0.37$, $df = 2$, $p\text{-value} = 0.83$



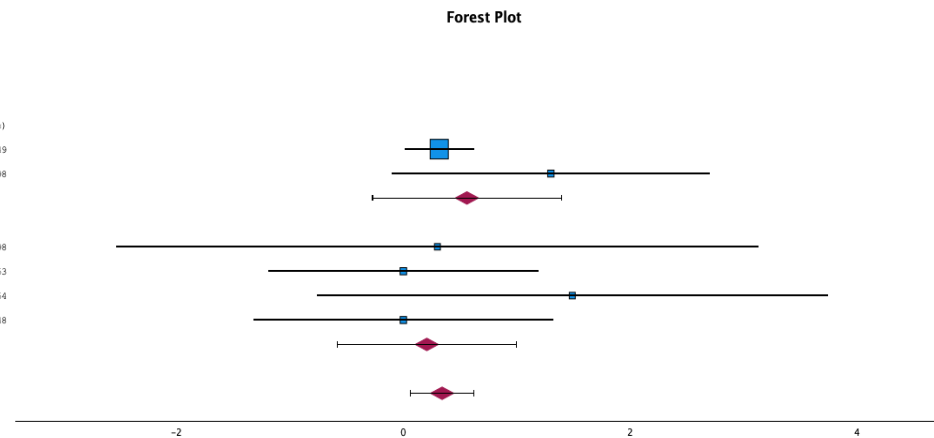
■ Effect size of each study
◆ Estimated overall effect size
| Estimated overall confidence interval

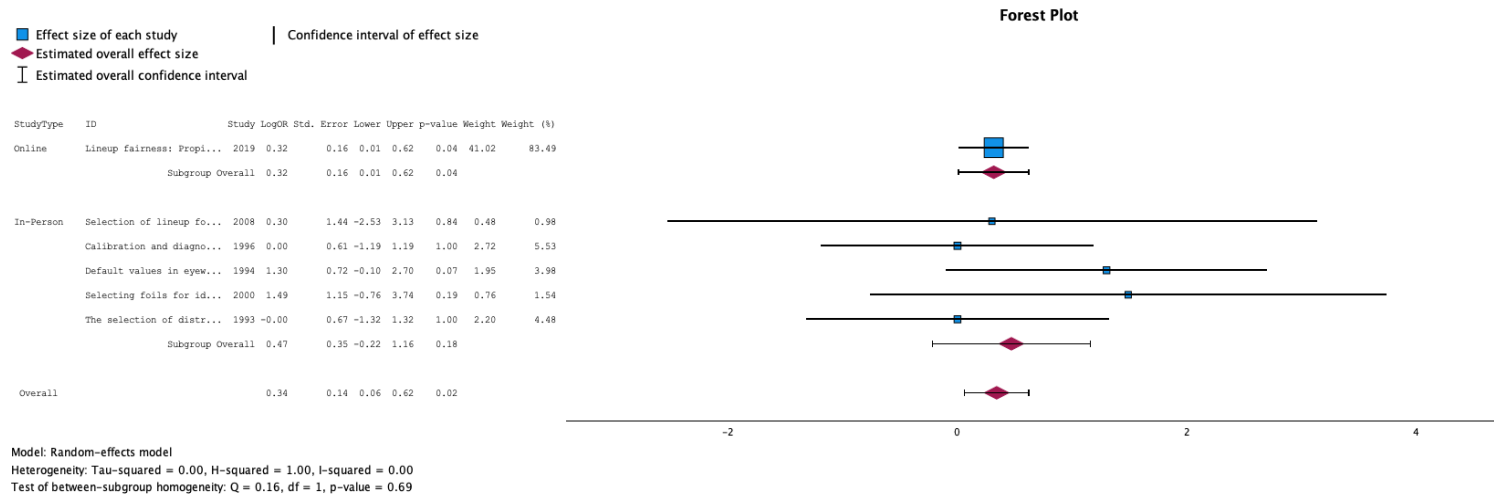
ParticipantType	ID	Study	LogOR	Std. Error	Lower	Upper	p-value	Weight	Weight (%)
Community Volunteer	Lineup fairness: Propi...	2019	0.32	0.16	0.01	0.62	0.04	41.02	83.49
	Default values in eyew...	1994	1.30	0.72	-0.10	2.70	0.07	1.95	3.98
	Subgroup Overall		0.56	0.42	-0.27	1.39	0.19		
University Student	Selection of lineup fo...	2008	0.30	1.44	-2.53	3.13	0.84	0.48	0.98
	Calibration and diagno...	1996	0.00	0.61	-1.19	1.19	1.00	2.72	5.53
	Selecting foils for id...	2000	1.49	1.15	-0.76	3.74	0.19	0.76	1.54
	The selection of distr...	1993	-0.00	0.67	-1.32	1.32	1.00	2.20	4.48
	Subgroup Overall		0.21	0.40	-0.58	1.00	0.61		
Overall			0.34	0.14	0.06	0.62	0.02		

Model: Random-effects model

Heterogeneity: Tau-squared = 0.00, H-squared = 1.00, I-squared = 0.00

Test of between-subgroup homogeneity: $Q = 0.36$, $df = 1$, $p\text{-value} = 0.55$

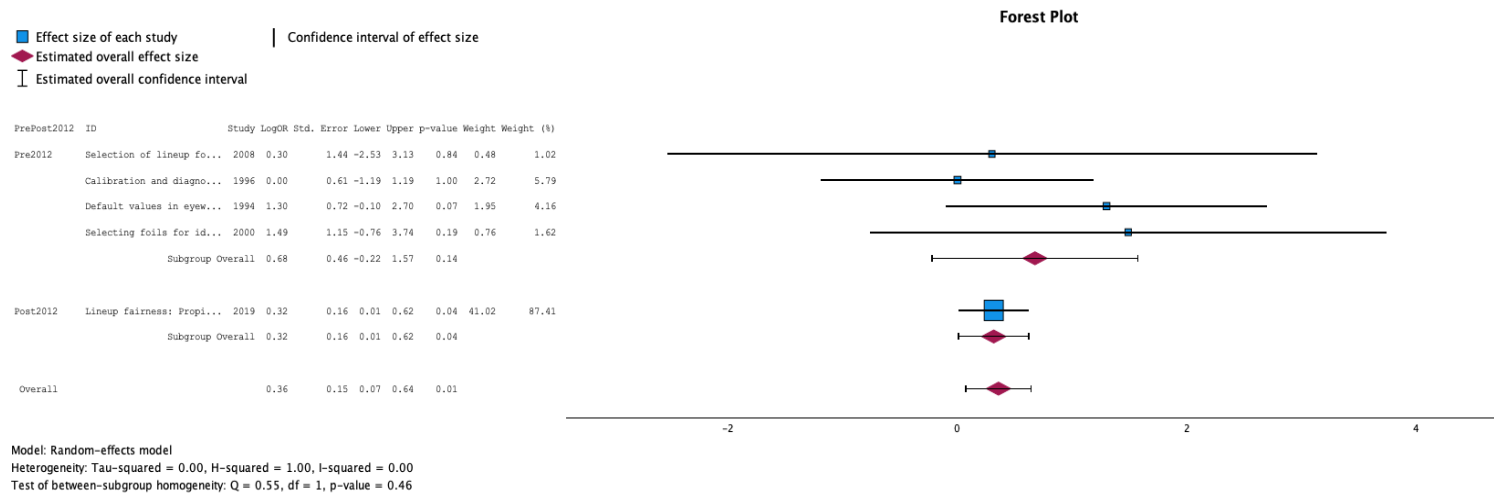




Note: Wells (1993) outlier is included in forest plot.

Figure 2

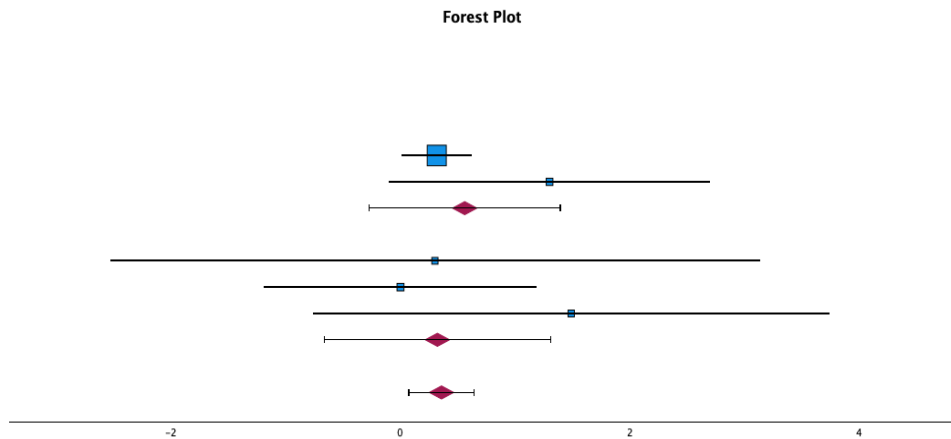
Forest Plot Depicting Moderator Analysis for Innocent Suspect IDs for Target Absent Lineups



■ Effect size of each study | Confidence interval of effect size
◆ Estimated overall effect size
┃ Estimated overall confidence interval

MTDtype	ID	Study	LogOR	Std. Error	Lower	Upper	p-value	Weight	Weight (%)
HybridMTD	Lineup fairness: Propi...	2019	0.32	0.16	0.01	0.62	0.04	41.02	87.41
	Default values in eyew...	1994	1.30	0.72	-0.10	2.70	0.07	1.95	4.16
	Subgroup Overall	0.56	0.42	-0.27	1.39	0.19			
PureMTD	Selection of lineup fo...	2008	0.30	1.44	-2.53	3.13	0.84	0.48	1.02
	Calibration and diagno...	1996	0.00	0.61	-1.19	1.19	1.00	2.72	5.79
	Selecting foils for id...	2000	1.49	1.15	-0.76	3.74	0.19	0.76	1.62
	Subgroup Overall	0.32	0.50	-0.66	1.31	0.52			
Overall			0.36	0.15	0.07	0.64	0.01		

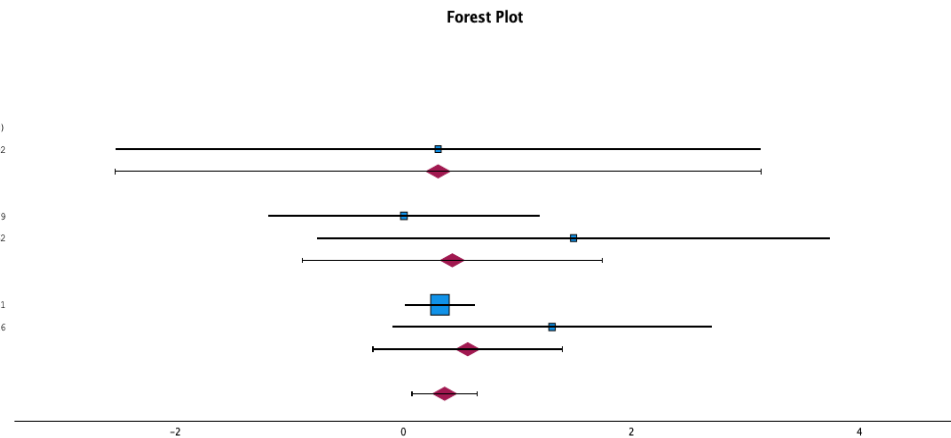
Model: Random-effects model
 Heterogeneity: Tau-squared = 0.00, H-squared = 1.00, I-squared = 0.00
 Test of between-subgroup homogeneity: Q = 0.13, df = 1, p-value = 0.72

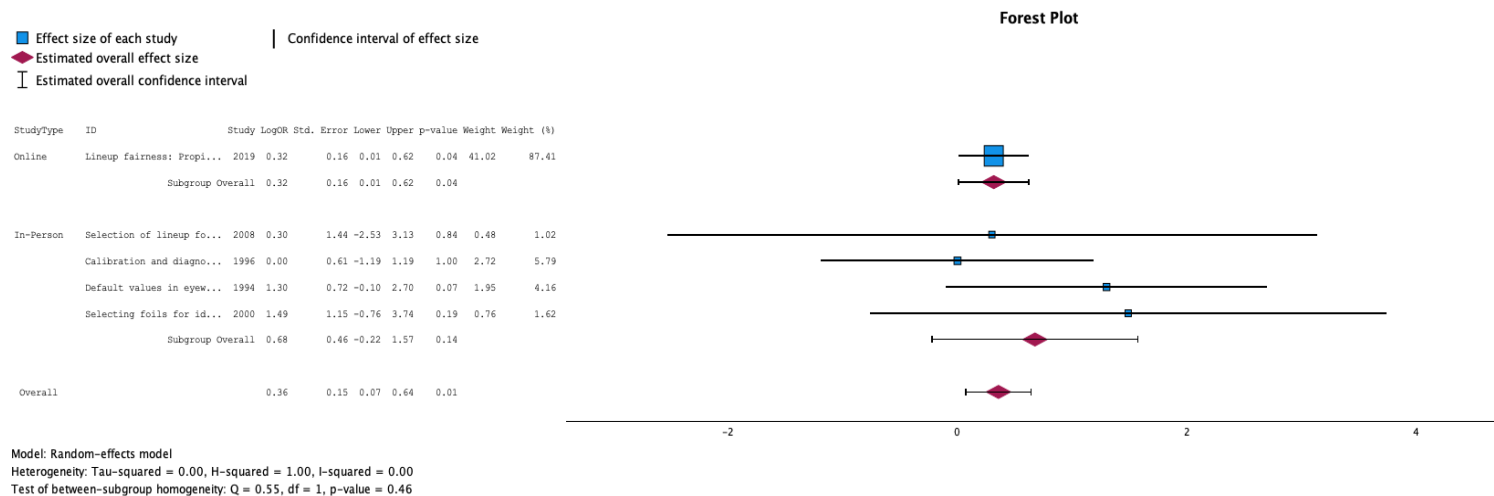
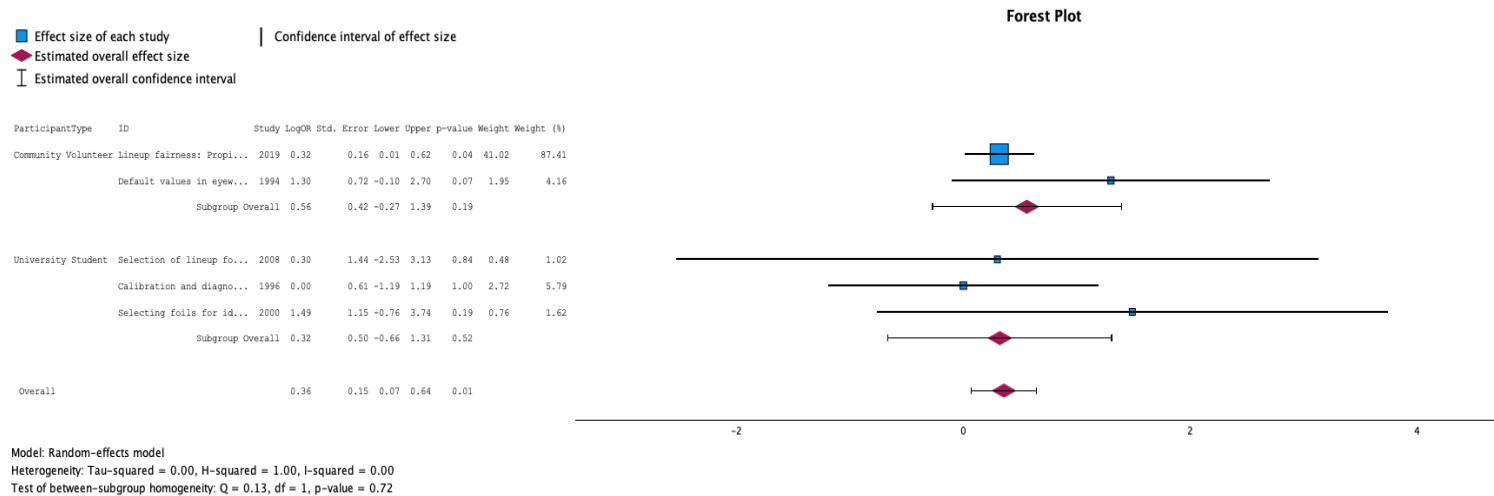


■ Effect size of each study | Confidence interval of effect size
◆ Estimated overall effect size
┃ Estimated overall confidence interval

SubsequentFilter	ID	Study	LogOR	Std. Error	Lower	Upper	p-value	Weight	Weight (%)
Unclear	Selection of lineup fo...	2008	0.30	1.44	-2.53	3.13	0.84	0.48	1.02
	Subgroup Overall	0.30	1.44	-2.53	3.13	0.84			
No	Calibration and diagno...	1996	0.00	0.61	-1.19	1.19	1.00	2.72	5.79
	Selecting foils for id...	2000	1.49	1.15	-0.76	3.74	0.19	0.76	1.62
	Subgroup Overall	0.42	0.67	-0.89	1.74	0.53			
Yes	Lineup fairness: Propi...	2019	0.32	0.16	0.01	0.62	0.04	41.02	87.41
	Default values in eyew...	1994	1.30	0.72	-0.10	2.70	0.07	1.95	4.16
	Subgroup Overall	0.56	0.42	-0.27	1.39	0.19			
Overall			0.36	0.15	0.07	0.64	0.01		

Model: Random-effects model
 Heterogeneity: Tau-squared = 0.00, H-squared = 1.00, I-squared = 0.00
 Test of between-subgroup homogeneity: Q = 0.05, df = 2, p-value = 0.97





Note: Forest plot displaying the moderator analysis results with Wells (1993) outlier removed.

Appendix H

Table 1

Test of Between-Subgroup Homogeneity for Filler IDs in Target Absent Lineups

Moderators Examined	Findings
Pre-2012 vs post-2012	$Q = 0.47, df = 1, p = 0.50$
MTD type (Hybrid MTD vs pure MTD)	$Q = 0.63, df = 1, p = 0.43$
Subsequent Filtering (Unclear, yes, no)	$Q = 1.88, df = 2, p = 0.39$
Participant Type (Community volunteer vs university student)	$Q = 0.63, df = 1, p = 0.43$
Study Type (Online vs in-person)	$Q = 0.47, df = 1, p = 0.50$

*Note: Wells (1993) outlier included in this table

Table 2

Test of Between-Subgroup Homogeneity for Filler IDs in Target Absent Lineups

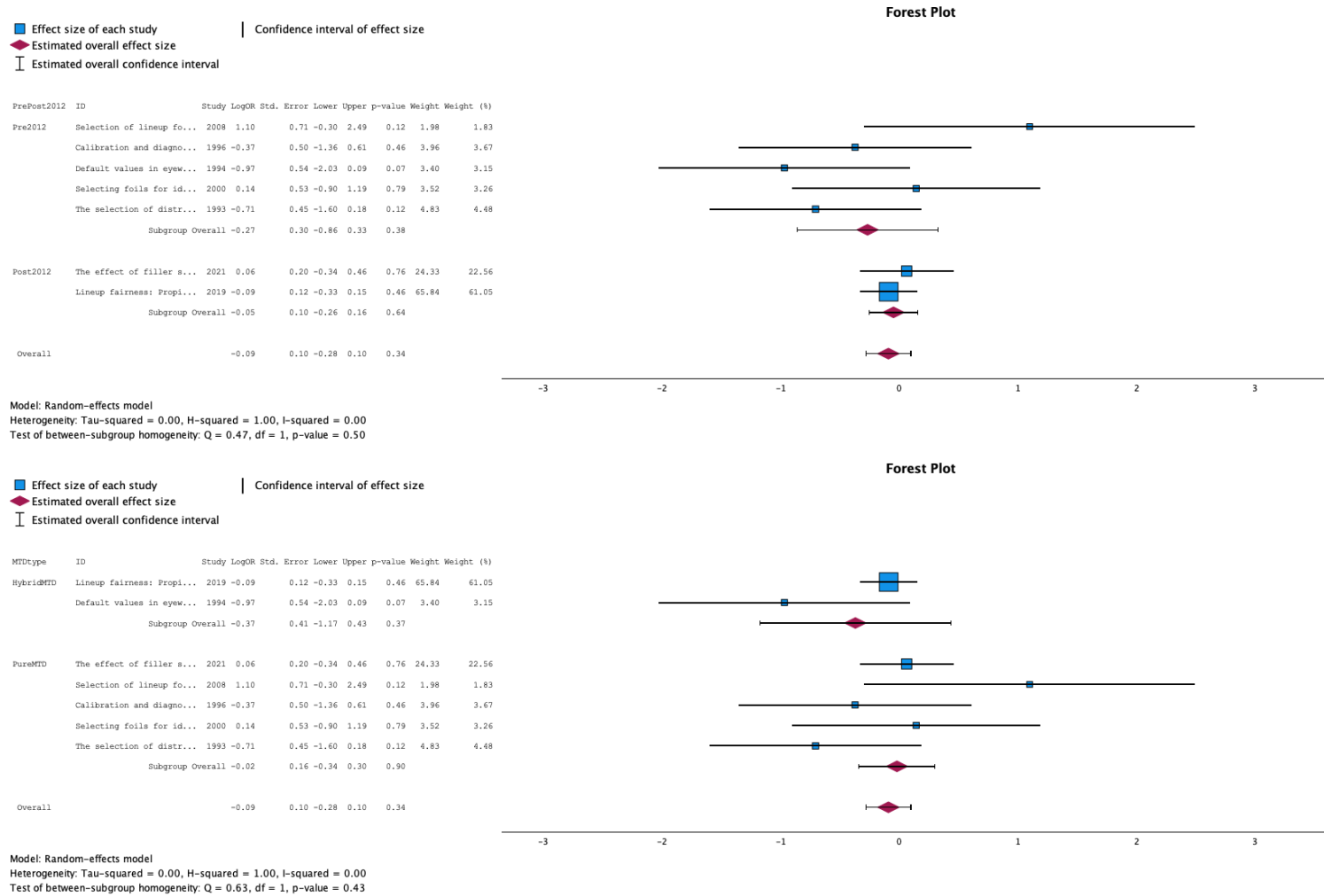
Moderators Examined	Findings
Pre-2012 vs post-2012	$Q = 0.02, df = 1, p = 0.88$
MTD type (Hybrid MTD vs pure MTD)	$Q = 1.03, df = 1, p = 0.31$
Subsequent Filtering (Unclear, yes, no)	$Q = 1.39, df = 2, p = 0.50$
Participant Type (Community volunteer vs university student)	$Q = 1.03, df = 1, p = 0.31$
Study Type (Online vs in-person)	$Q = 0.02, df = 1, p = 0.88$

*Note: Wells (1993) outlier not included in this table

Appendix I

Figure 1

Forest Plot Depicting Moderator Analyses for Filler IDs for Target Absent Lineups



■ Effect size of each study | Confidence interval of effect size
◆ Estimated overall effect size
┃ Estimated overall confidence interval

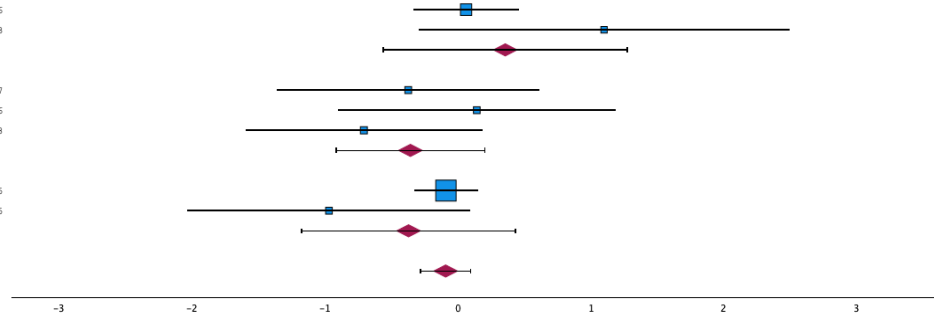
Subsequent/Filter ID	Study	LogOR	Std. Error	Lower	Upper	p-value	Weight	Weight (%)
Unclear	The effect of filler s...	2021	0.06	0.20	-0.34	0.46	0.76	24.33
	Selection of lineup fo...	2008	1.10	0.71	-0.30	2.49	0.12	1.98
	Subgroup Overall	0.36	0.47	-0.56	1.27	0.45		
No	Calibration and diagno...	1996	-0.37	0.50	-1.36	0.61	0.46	3.96
	Selecting foils for id...	2000	0.14	0.53	-0.90	1.19	0.79	3.52
	The selection of distr...	1993	-0.71	0.45	-1.60	0.18	0.12	4.83
	Subgroup Overall	-0.36	0.28	-0.92	0.20	0.21		
Yes	Lineup fairness: Propi...	2019	-0.09	0.12	-0.33	0.15	0.46	65.84
	Default values in eyew...	1994	-0.97	0.54	-2.03	0.09	0.07	3.40
	Subgroup Overall	-0.37	0.41	-1.17	0.43	0.37		
Overall			-0.09	0.10	-0.28	0.10	0.34	

Model: Random-effects model

Heterogeneity: Tau-squared = 0.00, H-squared = 1.00, I-squared = 0.00

Test of between-subgroup homogeneity: $Q = 1.88$, $df = 2$, $p\text{-value} = 0.39$

Forest Plot



■ Effect size of each study | Confidence interval of effect size
◆ Estimated overall effect size
┃ Estimated overall confidence interval

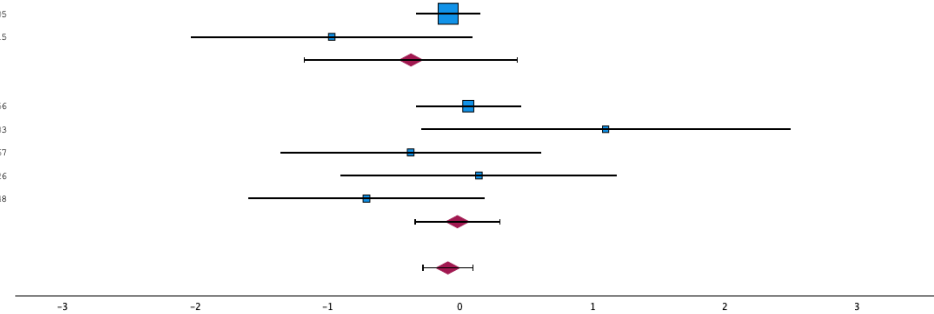
Participant/Type ID	Study	LogOR	Std. Error	Lower	Upper	p-value	Weight	Weight (%)
Community Volunteer	Lineup fairness: Propi...	2019	-0.09	0.12	-0.33	0.15	0.46	65.84
	Default values in eyew...	1994	-0.97	0.54	-2.03	0.09	0.07	3.40
	Subgroup Overall	-0.37	0.41	-1.17	0.43	0.37		
University Student	The effect of filler s...	2021	0.06	0.20	-0.34	0.46	0.76	24.33
	Selection of lineup fo...	2008	1.10	0.71	-0.30	2.49	0.12	1.98
	Calibration and diagno...	1996	-0.37	0.50	-1.36	0.61	0.46	3.96
	Selecting foils for id...	2000	0.14	0.53	-0.90	1.19	0.79	3.52
	The selection of distr...	1993	-0.71	0.45	-1.60	0.18	0.12	4.83
	Subgroup Overall	-0.02	0.16	-0.34	0.30	0.90		
Overall			-0.09	0.10	-0.28	0.10	0.34	

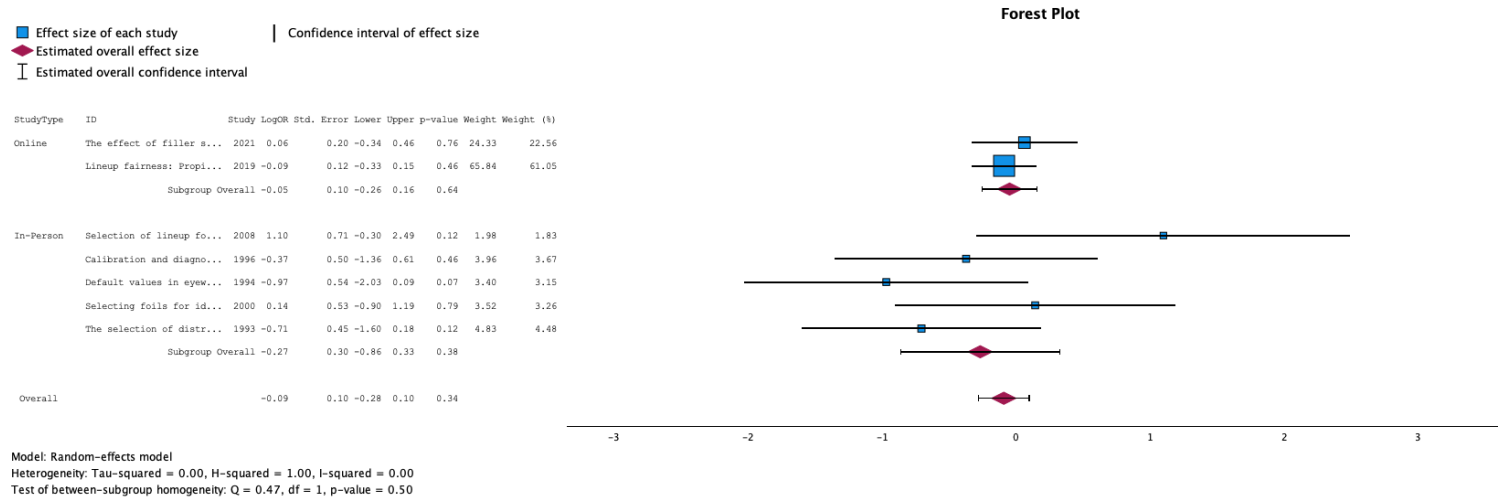
Model: Random-effects model

Heterogeneity: Tau-squared = 0.00, H-squared = 1.00, I-squared = 0.00

Test of between-subgroup homogeneity: $Q = 0.63$, $df = 1$, $p\text{-value} = 0.43$

Forest Plot

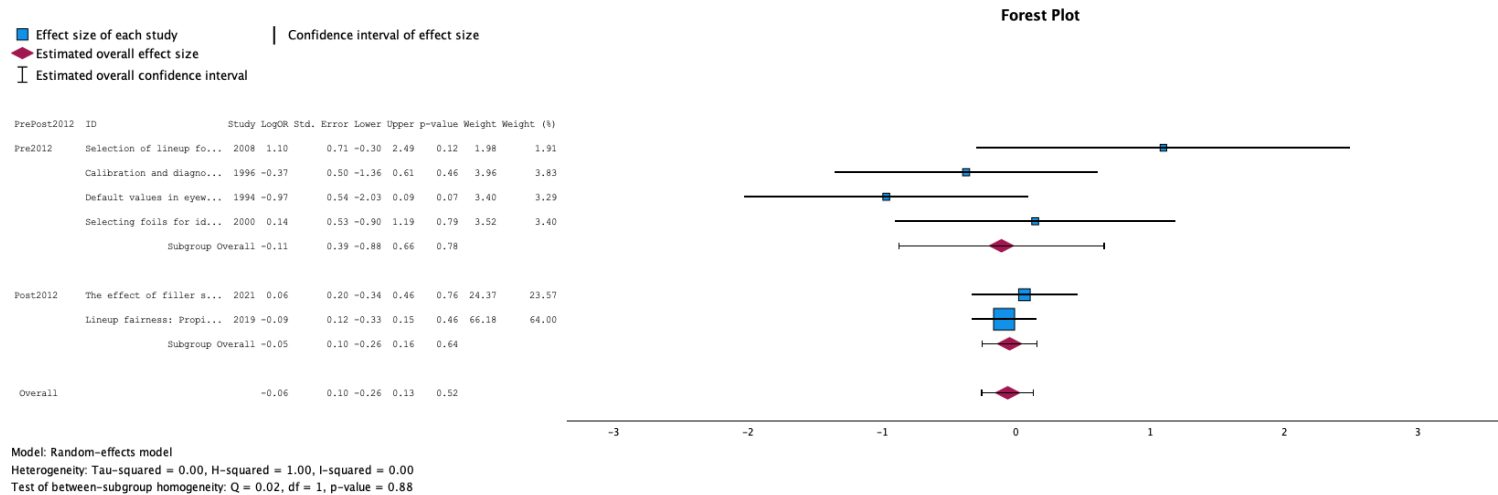




Note: Wells (1993) outlier is included in forest plot.

Figure 2

Forest Plot Depicting Moderator Analyses for Filler IDs for Target Absent Lineups



■ Effect size of each study
◆ Estimated overall effect size
| Estimated overall confidence interval

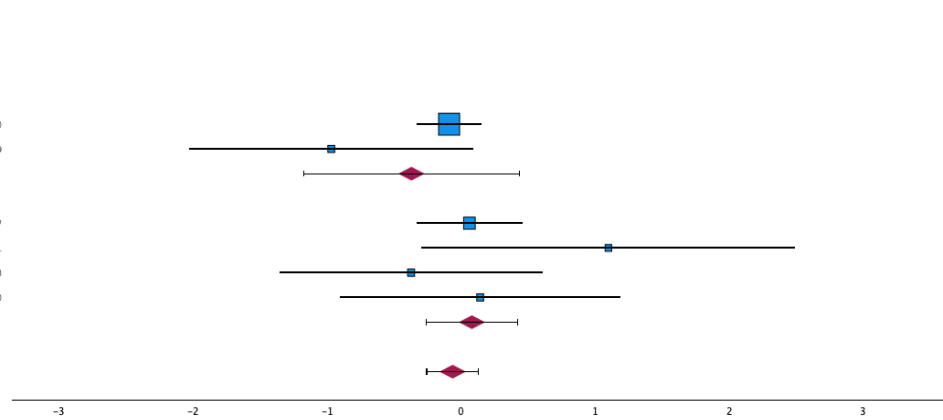
MDtype	ID	Study	LogOR	Std. Error	Lower	Upper	p-value	Weight	Weight (%)
HybridMD	Lineup fairness: Propi...	2019	-0.09	0.12	-0.33	0.15	0.46	66.18	64.00
	Default values in eyew...	1994	-0.97	0.54	-2.03	0.09	0.07	3.40	3.29
	Subgroup Overall		-0.37	0.41	-1.17	0.43	0.37		
PureMD	The effect of filler s...	2021	0.06	0.20	-0.34	0.46	0.76	24.37	23.57
	Selection of lineup fo...	2008	1.10	0.71	-0.30	2.49	0.12	1.98	1.91
	Calibration and diagno...	1996	-0.37	0.50	-1.36	0.61	0.46	3.96	3.83
	Selecting foils for id...	2000	0.14	0.53	-0.90	1.19	0.79	3.52	3.40
	Subgroup Overall		0.08	0.17	-0.26	0.42	0.65		
Overall			-0.06	0.10	-0.26	0.13	0.52		

Model: Random-effects model

Heterogeneity: Tau-squared = 0.00, H-squared = 1.00, I-squared = 0.00

Test of between-subgroup homogeneity: Q = 1.03, df = 1, p-value = 0.31

Forest Plot



■ Effect size of each study
◆ Estimated overall effect size
| Estimated overall confidence interval

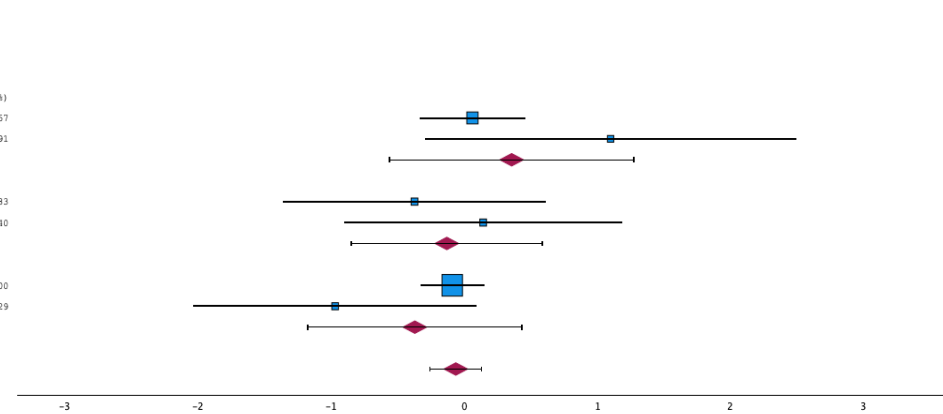
SubsequentFilter	ID	Study	LogOR	Std. Error	Lower	Upper	p-value	Weight	Weight (%)
Unclear	The effect of filler s...	2021	0.06	0.20	-0.34	0.46	0.76	24.37	23.57
	Selection of lineup fo...	2008	1.10	0.71	-0.30	2.49	0.12	1.98	1.91
	Subgroup Overall		0.36	0.47	-0.56	1.27	0.45		
No	Calibration and diagno...	1996	-0.37	0.50	-1.36	0.61	0.46	3.96	3.83
	Selecting foils for id...	2000	0.14	0.53	-0.90	1.19	0.79	3.52	3.40
	Subgroup Overall		-0.13	0.37	-0.85	0.59	0.72		
Yes	Lineup fairness: Propi...	2019	-0.09	0.12	-0.33	0.15	0.46	66.18	64.00
	Default values in eyew...	1994	-0.97	0.54	-2.03	0.09	0.07	3.40	3.29
	Subgroup Overall		-0.37	0.41	-1.17	0.43	0.37		
Overall			-0.06	0.10	-0.26	0.13	0.52		

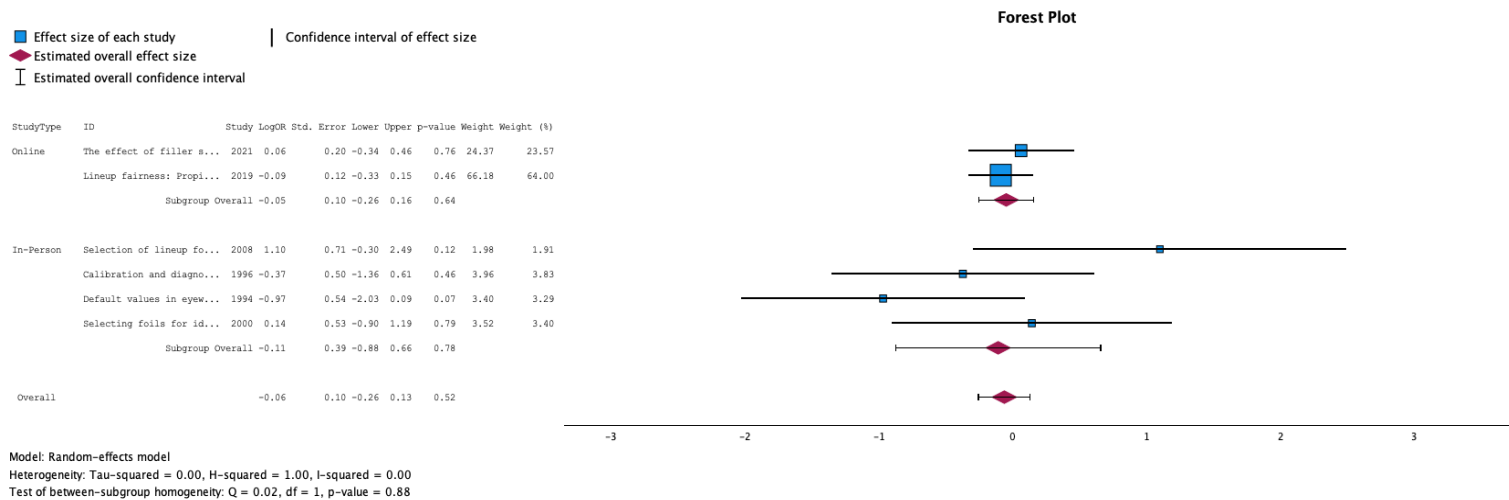
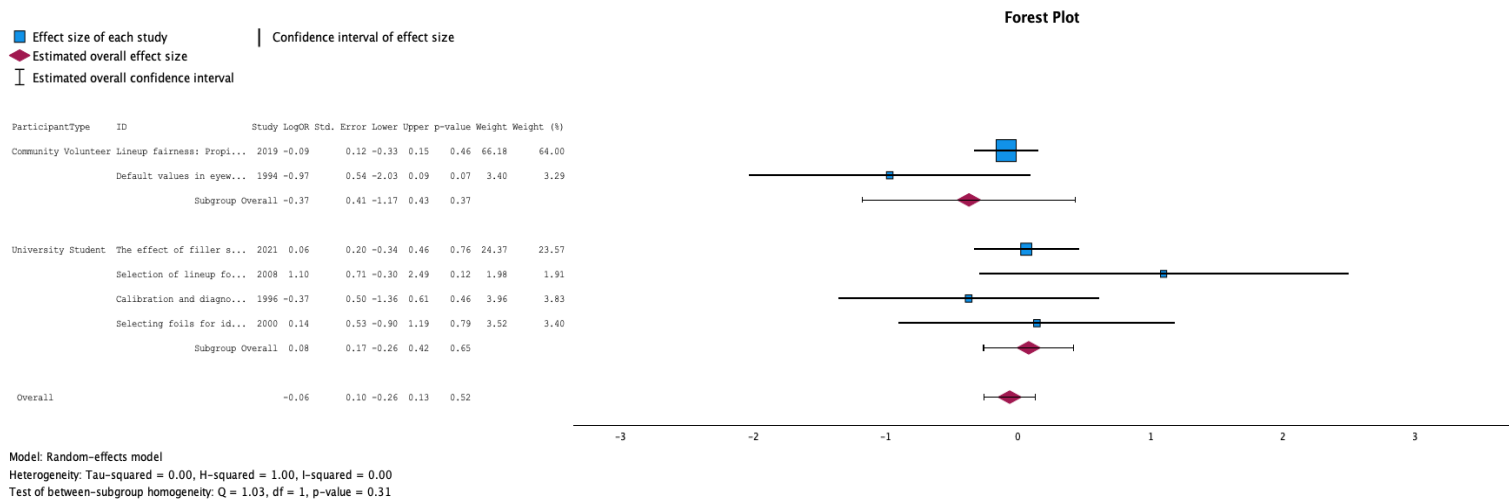
Model: Random-effects model

Heterogeneity: Tau-squared = 0.00, H-squared = 1.00, I-squared = 0.00

Test of between-subgroup homogeneity: Q = 1.39, df = 2, p-value = 0.50

Forest Plot





Note: Forest plot displaying the moderator analysis results with Wells (1993) outlier removed.

Appendix J

Table 1

Test of Between-Subgroup Homogeneity for Non-IDs in Target Absent Lineups

Moderators Examined	Findings
Pre-2012 vs post-2012	$Q = 0.10, df = 1, p = 0.75$
MTD type (Hybrid MTD vs pure MTD)	$Q = 0.04, df = 1, p = 0.85$
Subsequent Filtering (Unclear, yes, no)	$Q = 0.86, df = 2, p = 0.55$
Participant Type (Community volunteer vs university student)	$Q = 0.04, df = 1, p = 0.85$
Study Type (Online vs in-person)	$Q = 0.10, df = 1, p = 0.75$

*Note: Wells (1993) outlier included in this table

Table 2

Test of Between-Subgroup Homogeneity for Non-IDs in Target Absent Lineups

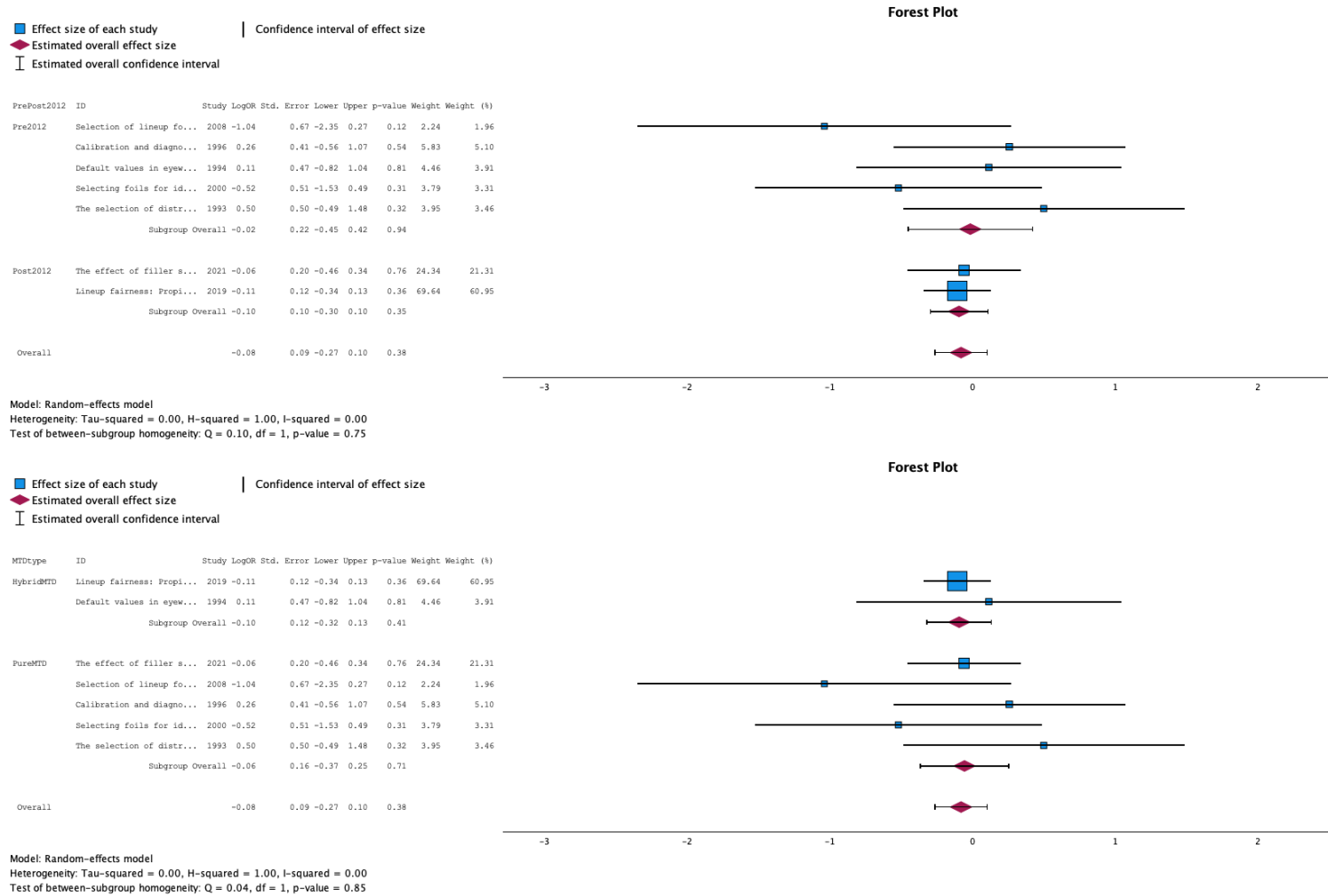
Moderators Examined	Findings
Pre-2012 vs post-2012	$Q = 0.04, df = 1, p = 0.84$
MTD type (Hybrid MTD vs pure MTD)	$Q = 0.01, df = 1, p = 0.91$
Subsequent Filtering (Unclear, yes, no)	$Q = 0.30, df = 2, p = 0.86$
Participant Type (Community volunteer vs university student)	$Q = 0.01, df = 1, p = 0.91$
Study Type (Online vs in-person)	$Q = 0.04, df = 1, p = 0.84$

*Note: Wells (1993) outlier not included in this table.

Appendix K

Figure 1

Forest Plot Depicting Moderator Analyses for Non-IDs for Target Absent Lineups



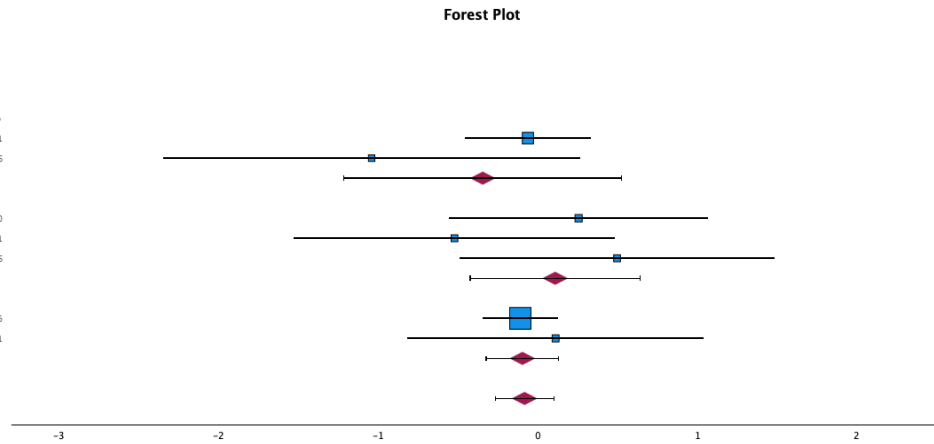
■ Effect size of each study
◆ Estimated overall effect size
| Estimated overall confidence interval

SubsequentFilter ID	Study LogOR Std. Error Lower Upper p-value Weight Weight (%)
Unclear	The effect of filler s... 2021 -0.06 0.20 -0.46 0.34 0.76 24.34 21.31
	Selection of lineup fo... 2008 -1.04 0.67 -2.35 0.27 0.12 2.24 1.96
	Subgroup Overall -0.34 0.44 -1.21 0.53 0.44
No	Calibration and diagno... 1996 0.26 0.41 -0.56 1.07 0.54 5.83 5.10
	Selecting foils for id... 2000 -0.52 0.51 -1.53 0.49 0.31 3.79 3.31
	The selection of distr... 1993 0.50 0.50 -0.49 1.48 0.32 3.95 3.46
	Subgroup Overall 0.11 0.27 -0.42 0.64 0.69
Yes	Lineup fairness: Propi... 2019 -0.11 0.12 -0.34 0.13 0.36 69.64 60.95
	Default values in eyew... 1994 0.11 0.47 -0.82 1.04 0.81 4.46 3.91
	Subgroup Overall -0.10 0.12 -0.32 0.13 0.41
Overall	-0.08 0.09 -0.27 0.10 0.38

Model: Random-effects model

Heterogeneity: Tau-squared = 0.00, H-squared = 1.00, I-squared = 0.00

Test of between-subgroup homogeneity: $Q = 0.86$, $df = 2$, $p\text{-value} = 0.65$



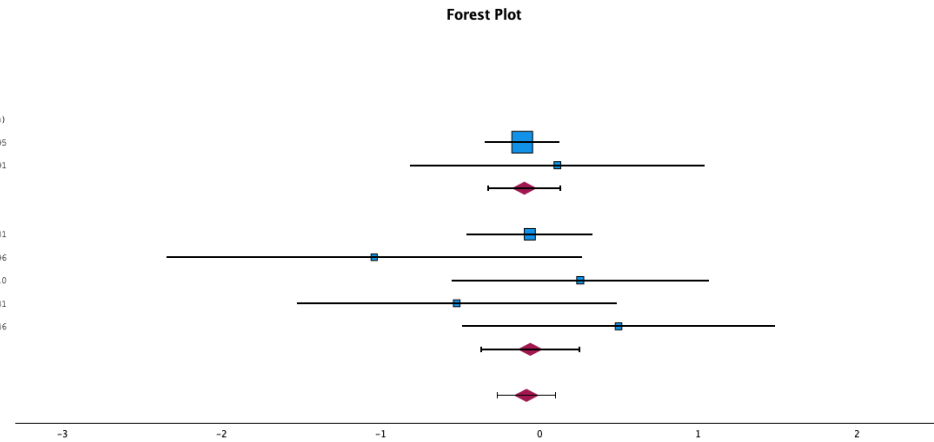
■ Effect size of each study
◆ Estimated overall effect size
| Estimated overall confidence interval

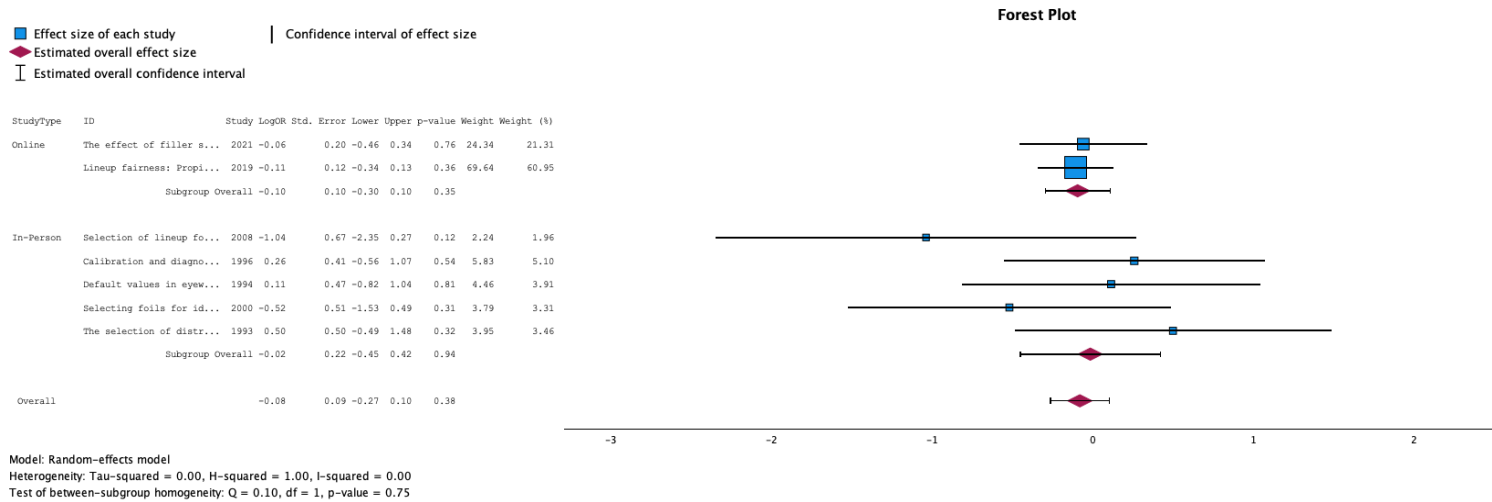
ParticipantType ID	Study LogOR Std. Error Lower Upper p-value Weight Weight (%)
Community Volunteer	Lineup fairness: Propi... 2019 -0.11 0.12 -0.34 0.13 0.36 69.64 60.95
	Default values in eyew... 1994 0.11 0.47 -0.82 1.04 0.81 4.46 3.91
	Subgroup Overall -0.10 0.12 -0.32 0.13 0.41
University Student	The effect of filler s... 2021 -0.06 0.20 -0.46 0.34 0.76 24.34 21.31
	Selection of lineup fo... 2008 -1.04 0.67 -2.35 0.27 0.12 2.24 1.96
	Calibration and diagno... 1996 0.26 0.41 -0.56 1.07 0.54 5.83 5.10
	Selecting foils for id... 2000 -0.52 0.51 -1.53 0.49 0.31 3.79 3.31
	The selection of distr... 1993 0.50 0.50 -0.49 1.48 0.32 3.95 3.46
	Subgroup Overall -0.06 0.16 -0.37 0.25 0.71
Overall	-0.08 0.09 -0.27 0.10 0.38

Model: Random-effects model

Heterogeneity: Tau-squared = 0.00, H-squared = 1.00, I-squared = 0.00

Test of between-subgroup homogeneity: $Q = 0.04$, $df = 1$, $p\text{-value} = 0.85$

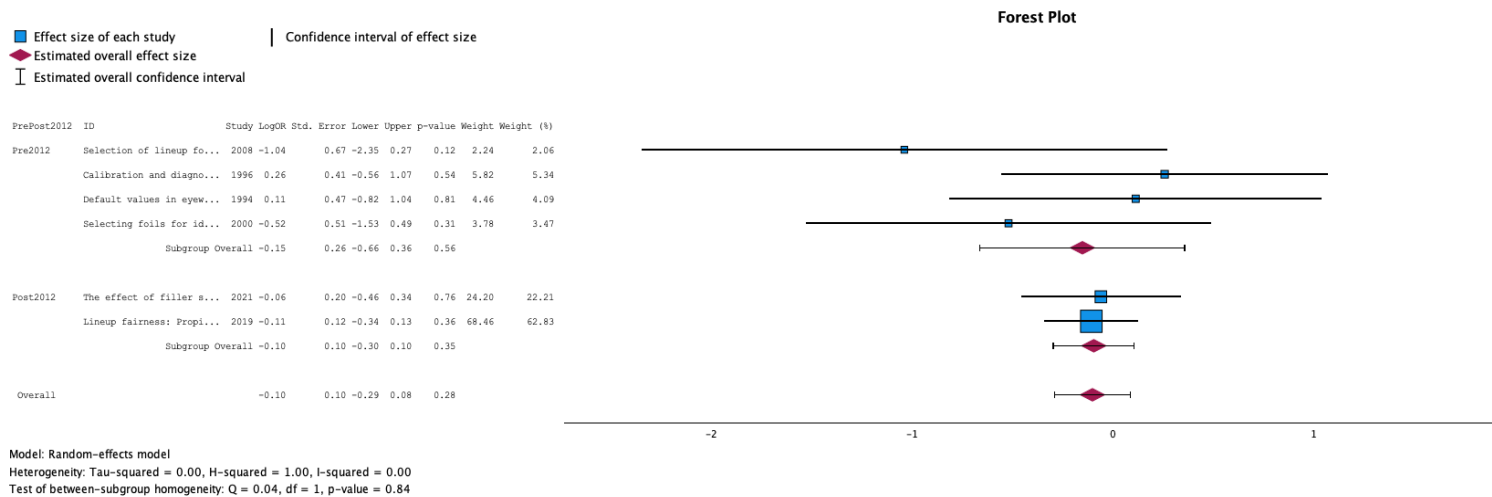




Note: Wells (1993) outlier is included in forest plot.

Figure 2

Forest Plot Depicting Moderator Analysis for Non-IDs for Target Absent Lineups

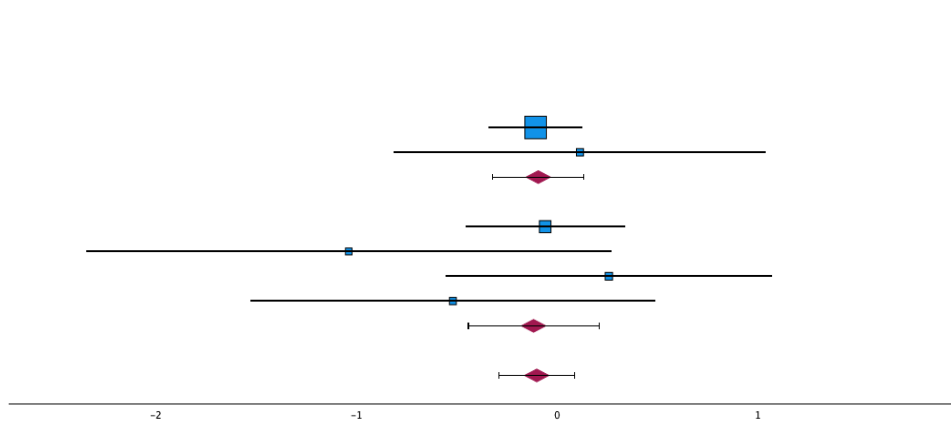


■ Effect size of each study | Confidence interval of effect size
◆ Estimated overall effect size
└ Estimated overall confidence interval

MDtype	ID	Study	LogOR	Std. Error	Lower	Upper	p-value	Weight	Weight (%)
HybridMD	Lineup fairness: Propi...	2019	-0.11	0.12	-0.34	0.13	0.36	68.46	62.83
	Default values in eyew...	1994	0.11	0.47	-0.82	1.04	0.81	4.46	4.09
	Subgroup Overall	-0.10	0.12	-0.32	0.13	0.41			
PureMD	The effect of filler s...	2021	-0.06	0.20	-0.46	0.34	0.76	24.20	22.21
	Selection of lineup fo...	2008	-1.04	0.67	-2.35	0.27	0.12	2.24	2.06
	Calibration and diagno...	1996	0.26	0.41	-0.56	1.07	0.54	5.82	5.34
	Selecting foils for id...	2000	-0.52	0.51	-1.53	0.49	0.31	3.78	3.47
	Subgroup Overall	-0.12	0.17	-0.44	0.21	0.47			
Overall		-0.10	0.10	-0.29	0.08	0.28			

Model: Random-effects model
Heterogeneity: Tau-squared = 0.00, H-squared = 1.00, I-squared = 0.00
Test of between-subgroup homogeneity: Q = 0.01, df = 1, p-value = 0.91

Forest Plot

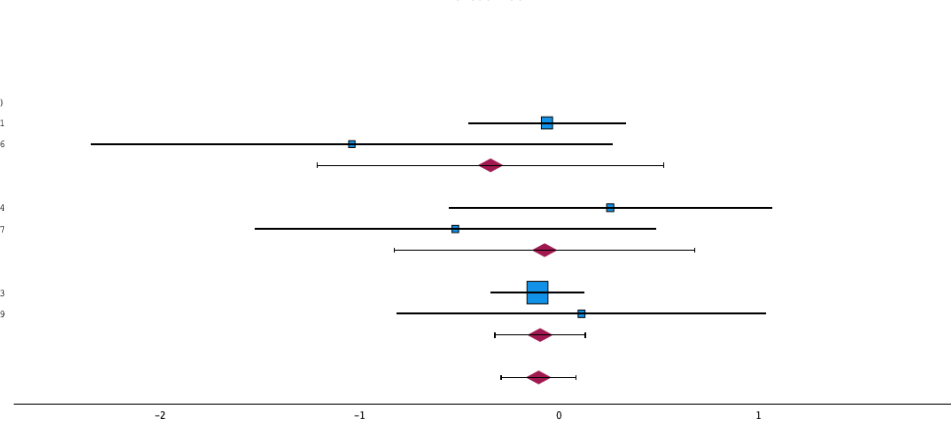


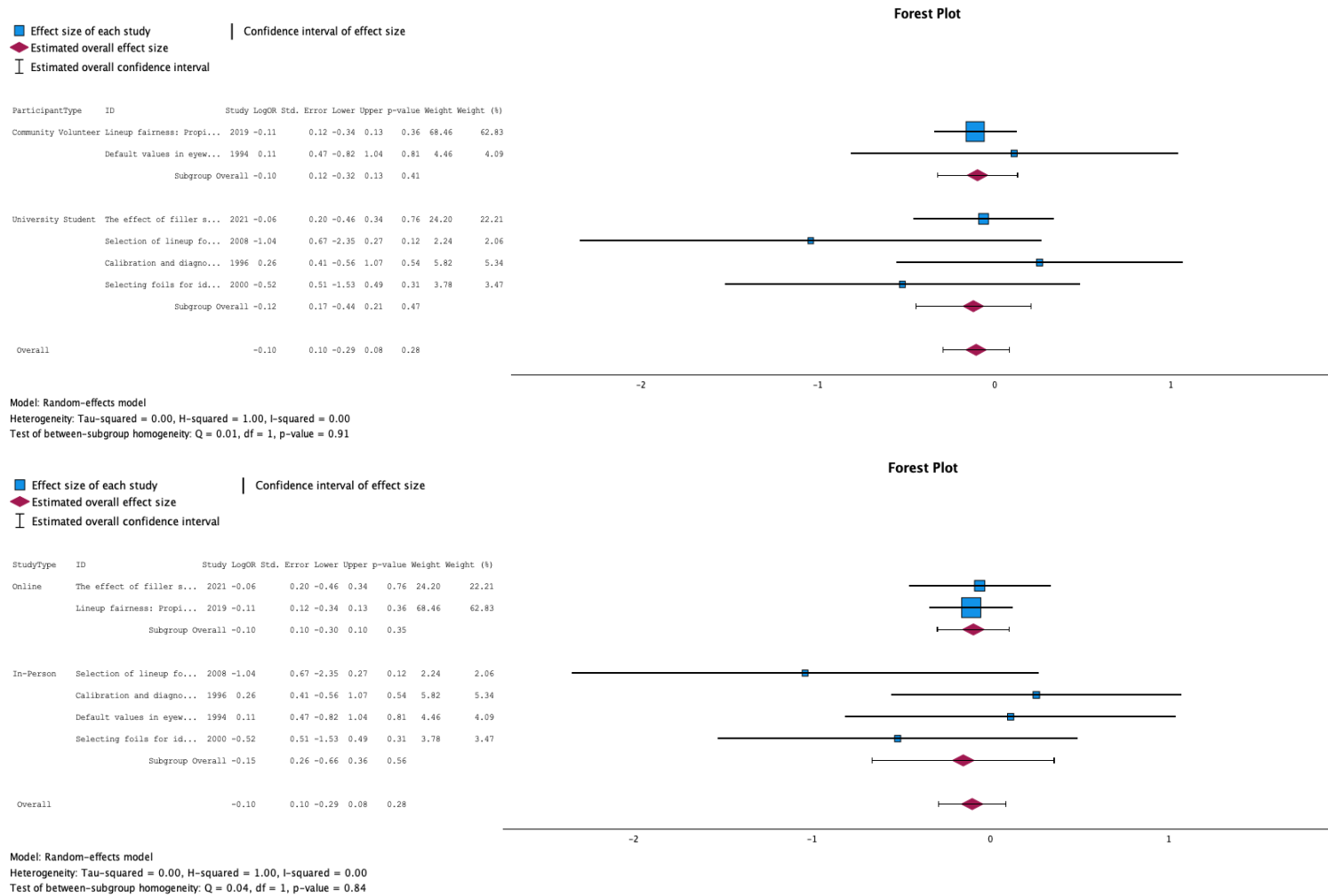
■ Effect size of each study | Confidence interval of effect size
◆ Estimated overall effect size
└ Estimated overall confidence interval

SubsequentFilter	ID	Study	LogOR	Std. Error	Lower	Upper	p-value	Weight	Weight (%)
Unclear	The effect of filler s...	2021	-0.06	0.20	-0.46	0.34	0.76	24.20	22.21
	Selection of lineup fo...	2008	-1.04	0.67	-2.35	0.27	0.12	2.24	2.06
	Subgroup Overall	-0.34	0.44	-1.21	0.53	0.44			
No	Calibration and diagno...	1996	0.26	0.41	-0.56	1.07	0.54	5.82	5.34
	Selecting foils for id...	2000	-0.52	0.51	-1.53	0.49	0.31	3.78	3.47
	Subgroup Overall	-0.07	0.38	-0.83	0.68	0.85			
Yes	Lineup fairness: Propi...	2019	-0.11	0.12	-0.34	0.13	0.36	68.46	62.83
	Default values in eyew...	1994	0.11	0.47	-0.82	1.04	0.81	4.46	4.09
	Subgroup Overall	-0.10	0.12	-0.32	0.13	0.41			
Overall		-0.10	0.10	-0.29	0.08	0.28			

Model: Random-effects model
Heterogeneity: Tau-squared = 0.00, H-squared = 1.00, I-squared = 0.00
Test of between-subgroup homogeneity: Q = 0.30, df = 2, p-value = 0.86

Forest Plot





Note: Forest plot displaying the moderator analysis results with Wells (1993) outlier removed.

Appendix L

Modified Missing Persons SAI, from Section 3 of Missing Persons SAI

Please envision the perpetrator seen in the video. Hold the image of this person in your mind, then read through the list of memory cues below to see if they help you remember information that we will find useful.

- Height / Weight / Build (approximate or exact)
- Facial appearance (shape of eyes, ears, mouth, nose, etc.)
- Complexion: (clear skin, freckles, acne, ruddy, etc.)
- Teeth: (condition, gaps, chips, gold teeth, etc.)
- Hair: Style / Colour / Length
- Facial hair / body hair: Style / Colour
- Nails: (manicured, long, short, bitten, nail-polish, etc.)
- Any identifying marks: Scars / Tattoos / Piercings / etc. (give locations)
- Distinguishing physical characteristics / deformities (give details)

Please now provide a detailed physical description of the perpetrator, be as complete and accurate as possible.

Please do not guess on any details which you are unsure.

Appendix M

Default Value Checklist

Please select the ethnicity of the individual from the video.

- ☐ Caucasian/White
- ☐ Black African
- ☐ Black Other
- ☐ Chinese
- ☐ Filipino
- ☐ Indian
- ☐ Irish Traveler
- ☐ Mixed
- ☐ Roma
- ☐ Hispanic/Latino
- ☐ Other
- ☐ I don't know

Please select the sex of the individual in the video.

- ☐ Male
- ☐ Female
- ☐ Other
- ☐ I don't know

Please select approximate age range for individual in the video.

- ☐ 18-24
- ☐ 25-34
- ☐ 35-44
- ☐ 45-54
- ☐ 55-64
- ☐ 65+

Please select approximate body type for individual in the video. (Select all that apply)

- ☐ Ectomorph - lean and slender build
- ☐ Endomorph - heavy-set build
- ☐ Mesomorph - athletic build
- ☐ Combination
- ☐ Other
- ☐ I don't know

Appendix N

Lineup Sheet Provided to Lineup Constructors

- Step 1: Assign any **one** of the boxes below (1-6) to the suspect by writing “suspect” in the chosen box.
- Step 2: Select 5 individuals from the database which BEST match the culprit image/description provided and write their corresponding photo IDs in the remaining boxes in the order you choose.

1
PhotoID: _____

2
PhotoID: _____

3
PhotoID: _____

4
PhotoID: _____

5
PhotoID: _____

6
PhotoID: _____

Appendix O

Difficulty Rating Scale and Decision-Making Process Statements

How easy or difficult did you find it to choose the best match to the description?

Very difficult	Difficult	Neither easy nor difficult	Easy	Very easy
----------------	-----------	-------------------------------	------	-----------

For each statement below, please select Yes if it describes your decision process, or No if it does not.

The face I chose just "popped out" at me

Yes	No
-----	----

I compared the photos to each other in order to narrow the choices

Yes	No
-----	----

I first eliminated the ones that didn't seem to fit, then chose among the rest

Yes	No
-----	----

Upon reading the description, I formed an image of the person in my head. I then selected the person who best matched that description.

Yes	No
-----	----

I based the judgment on specific facial features

Yes	No
-----	----

Next

Appendix P

PI-20 Index

Instructions: Below are a series of statements regarding facial recognition. You will be asked to rate your level of agreement on a 5-point scale, ranging from 1 (Strongly disagree) to 5 (Strongly agree). Please take your time reading through these statements and note that there is no right or wrong answer.

1. My face recognition ability is worse than most people

Strongly Disagree	1	2	3	4	5	Strongly Agree
----------------------	---	---	---	---	---	----------------

2. I have always had a bad memory for faces

Strongly Disagree	1	2	3	4	5	Strongly Agree
----------------------	---	---	---	---	---	----------------

3. I find it notably easier to recognize people who have distinctive facial features

Strongly Disagree	1	2	3	4	5	Strongly Agree
----------------------	---	---	---	---	---	----------------

4. I often mistake people I have met before for strangers

Strongly Disagree	1	2	3	4	5	Strongly Agree
----------------------	---	---	---	---	---	----------------

5. When I was at school, I struggled to recognize my classmates

Strongly Disagree	1	2	3	4	5	Strongly Agree
----------------------	---	---	---	---	---	----------------

6. When people change their hairstyle or wear hats, I have problems recognizing them

Strongly Disagree	1	2	3	4	5	Strongly Agree
----------------------	---	---	---	---	---	----------------

7. I sometimes have to warn new people I meet that I am 'bad with faces'

Strongly Disagree	1	2	3	4	5	Strongly Agree
----------------------	---	---	---	---	---	----------------

8. I find it easy to picture individual faces in my mind

Strongly Disagree	1	2	3	4	5	Strongly Agree
----------------------	---	---	---	---	---	----------------

9. I am better than most people at putting a 'name to a face'

Strongly Disagree	1	2	3	4	5	Strongly Agree
----------------------	---	---	---	---	---	----------------

10. Without hearing people's voices, I struggle to recognize them

Strongly Disagree	1	2	3	4	5	Strongly Agree
----------------------	---	---	---	---	---	----------------

11. Anxiety about face recognition has led me to avoid certain social or professional situations

Strongly Disagree	1	2	3	4	5	Strongly Agree
----------------------	---	---	---	---	---	----------------

12. I have to try harder than other people to memorize faces

Strongly Disagree	1	2	3	4	5	Strongly Agree
----------------------	---	---	---	---	---	----------------

13. I am very confident in my ability to recognize myself in photographs

Strongly Disagree	1	2	3	4	5	Strongly Agree
----------------------	---	---	---	---	---	----------------

14. I sometimes find movies hard to follow because of difficulties recognizing characters

Strongly Disagree	1	2	3	4	5	Strongly Agree
----------------------	---	---	---	---	---	----------------

15. My friends and family think I have bad face recognition or bad face memory

Strongly Disagree	1	2	3	4	5	Strongly Agree
----------------------	---	---	---	---	---	----------------

15. My friends and family think I have bad face recognition or bad face memory

Strongly Disagree	1	2	3	4	5	Strongly Agree
----------------------	---	---	---	---	---	----------------

16. I feel like I frequently offend people by not recognizing who they are

Strongly Disagree	1	2	3	4	5	Strongly Agree
----------------------	---	---	---	---	---	----------------

17. It is easy for me to recognize individuals in situations that require people to wear similar clothes (e.g. suits, uniforms and swimwear)

Strongly Disagree	1	2	3	4	5	Strongly Agree
----------------------	---	---	---	---	---	----------------

18. At family gatherings, I sometimes confuse individual family members

Strongly Disagree	1	2	3	4	5	Strongly Agree
----------------------	---	---	---	---	---	----------------

19. I find it easy to recognize celebrities in 'before-they-were-famous' photos, even if they have changed considerably

Strongly Disagree	1	2	3	4	5	Strongly Agree
----------------------	---	---	---	---	---	----------------

20. It is hard to recognize familiar people when I meet them out of context (e.g. meeting a work colleague unexpectedly while shopping)

Strongly Disagree	1	2	3	4	5	Strongly Agree
----------------------	---	---	---	---	---	----------------

Next ➔

Appendix Q

Example of lineup identification task

Below is a lineup of individuals who *may or may not* have been in the previously viewed video. Please take a look at each individual before making your identification. Once you have made your decision, please select the option with the corresponding number 1-6.

Please note the perpetrator *may be absent* from this lineup, in which case there is a "None of these / Individual is not present" option.



1



2



3



4



5



6

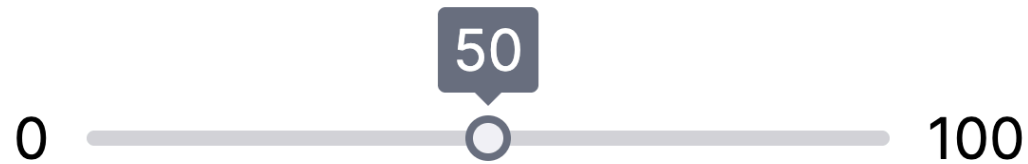
- ☐ 1
- ☐ 2
- ☐ 3
- ☐ 4
- ☐ 5
- ☐ 6
- ☐ None of these / Individual is not present

[Next](#)

Appendix R

An example of the confidence rating scale

How confident are you in your identification decision on a scale of 0-100, with 0 being not confident and 100 being extremely confident? Please move the dial in either direction to rate confidence before hitting the 'Next' button.



Next