Identity crisis in identification evidence: Similarity judgments as an alternative to identification decisions

by

Micaiah Zwartz

BA (Psychology), Grad Dip (Psychology)

Submitted in partial fulfillment of the requirements for the degree of

Doctor of Psychology (Forensic)

Deakin University

February 2016
I am the author of the thesis entitled

Identity crisis in identification evidence: Similarity judgments as an alternative to identification decisions

submitted for the degree of

Doctor of Psychology (Forensic)

This thesis may be made available for consultation, loan and limited copying in accordance with the Copyright Act 1968.

'I certify that I am the student named below and that the information provided in the form is correct'

Full Name: Micaiah Timothy Peter Zwartz

Signed: [Signature Redacted by Library]

Date: 27 January 2016
Identity Crisis in Identification Evidence

DEAKIN UNIVERSITY
ACCESS TO THESIS - A

I certify the following about the thesis entitled:

Identity crisis in identification evidence: Similarity judgments as an alternative to identification decisions

submitted for the degree of Doctor of Psychology (Forensic)

a. I am the creator of all or part of the whole work(s) (including content and layout) and that where reference is made to the work of others, due acknowledgment is given.

b. The work(s) are not in any way a violation or infringement of any copyright, trademark, patent, or other rights whatsoever of any person.

c. That if the work(s) have been commissioned, sponsored or supported by any organisation, I have fulfilled all of the obligations required by such contract or agreement.

d. That any material in the thesis which has been accepted for a degree or diploma by any university or institution is identified in the text.

e. All research integrity requirements have been complied with.

'I certify that I am the student named below and that the information provided in the form is correct'

Full Name: Micaiah Timothy Peter Zwartz

Signed: Signature Redacted by Library

Date: 27 January 2016
Acknowledgements

A number of individuals have made the completion of this thesis possible. I am deeply indebted to them. Thank you to my primary supervisor Don Thomson for his patience, wisdom, effort, and grace in the face of ongoing defeat (it is worth noting that the Collingwood Football Club have not beaten the Hawthorn Football Club once during the completion of this doctorate – may this remain ever so). Thank you to my secondary supervisor Adrian Scott for the major role he played in the setting up of my research project, as well as for his ongoing input and infectious enthusiasm. Thank you to my fellow D-psych colleagues for making the journey more gratifying than it might otherwise have been. Thank you to the Deakin University staff, particularly those in the forensic program, for treating us more like colleagues than students, and for passing on their passion. Thank you to my family, and in particular my parents, for their support and for modeling assiduous service. Thanks to my siblings will be provided upon their acknowledging that they will call me Doctor Mic and not Proctor Dick as threatened. Most importantly, special thanks to Una for making the completion of this doctorate possible, for the many sacrifices you have made, and for your ongoing and unfailing love and support.

In memory of Sam and Tristan.
# Table of Contents

Acknowledgments

Table of Contents

List of Appendices

List of Tables and Figures

Abstract

Overview of Thesis

**CHAPTER 1: Identification evidence in the legal context: Definitions and difficulties**

1.1 Definition of Identification evidence: The Uniform Evidence Acts

1.2 Identification parades and photographic lineups

1.3 Positive identification and resemblance evidence

1.4 Problems with the distinction between positive identification and resemblance evidence

   i. Problems relating to the consequences of the distinction between positive identification and resemblance evidence

   ii. Problems with the conception, collection and presentation of identification evidence

1.5 Evidence of resemblance is all a witness can give

**CHAPTER 2: Current practice in identification procedures: Research and applications**

2.1 The selection of foils

2.2 Simultaneous and sequential presentation

2.3 Two factors affecting identification decisions: Discrimination and response criterion

2.4 Factors affecting discrimination

   i. Own-race bias

   ii. Own-gender bias

   iii. Other factors affecting discrimination

2.5 Factors affecting the response criterion

   i. Lineup procedure (SIML and SEQL)

   ii. The relationship between discrimination and response criterion is blurred

   iii. Similarity processes

   iv. Lineup instructions

   v. The position of the target: Order effects and counterbalancing
i. Differentiation and Match: A method for exploring the relationship between similarity ratings and identifications..........................60
ii. Graphic representation of the eight lineups......................................65

7.2 Discussion (Studies 1 & 2) .................................................................65
   i. Are similarity ratings related to identification patterns? ...............66
   ii. Under what conditions is the relationship strongest? ...............66
      a. Reasons for a disconnect between similarity and identity ......68
   iii. Do similarity ratings provide a useful and viable alternative
to identifications? ........................................................................69
      a. Did ratings distinguish between targets and foils, and targets
         and target replacements? ..................................................69
      b. Were ratings diagnostic of accuracy? ..................................69
      c. Advantages of using ratings.............................................70
   iv. How do similarity ratings differ across experimental conditions? .....71
   v. Limitations and recommendations.........................................71
   vi. Summary of findings............................................................73

CHAPTER 8: Study 3........................................................................73

8.1 Method.........................................................................................74
   i. Participants..............................................................................74
   ii. Design....................................................................................74
   iii. Materials.............................................................................76
   iv. Procedure.............................................................................77

8.2 Results..........................................................................................78
   i. The impact of condition: Visual versus memory task..............81
   ii. The impact of lineup procedure: Simultaneous and sequential ....82
   iii. The impact of target presence/absence...............................83
   iv. The impact of order of presentation: Early versus late ..........85
   v. The impact of participant gender...........................................87
   vi. The impact of participant ethnicity........................................88
   vii. Discrepancy between the highest and second highest ratings as an
        indicator of accuracy.........................................................89

CHAPTER 9: Study 4........................................................................94

9.1 Method.........................................................................................94
   i. Participants..............................................................................94
   ii. Design....................................................................................94
   iii. Materials.............................................................................95
   iv. Procedure.............................................................................95
9.2 Results

i. The impact of lineup procedure

ii. The impact of order of presentation

iii. The impact of gender

iv. The impact of ethnicity

CHAPTER 10: Comparing Study 3 & Study 4

10.1 Results (Studies 3 & 4)

i. A problem with the SEQL identification procedure: The target may not be seen

ii. Differentiation and Match

a. Visual and memory task

b. SIML & SEQL procedure

c. Target present compared to target absent

iii. Graphic representation of the 8 lineups

10.2 Discussion (Studies 3 & 4)

i. Further evidence for similarity judgments underlying identifications

ii. Are ratings more accurate than identifications?

iii. SIML & SEQL procedure: Ratings

a. Reasons for the SEQL outperforming the SIML with regards to ratings

iv. SIML & SEQL procedure: Identifications

v. Similarity is subjective

vi. Visual versus memory conditions

vii. Order effects: Ratings

viii. Order effects: Identifications

ix. Discrepancy as a measure of accuracy

x. Target present and target absent ratings

xi. The impact of gender and ethnicity

CHAPTER 11: General Discussion

11.1 The existence of a relationship between similarity ratings and identifications

Preliminary evidence obtained

11.2 Are similarity ratings diagnostic of accuracy?

11.3 Reasons why SIML ratings were less accurate than identifications

i. Response criterion explanation – threshold responding was better captured by ratings

ii. Identifications may have relied more on global/automatic processing
whereas ratings encouraged local/conscious processing………123

iii. Rating every lineup member may have increased memory
    interference…………………………………………………………125

iv. Identifications do not necessarily depend on similarity…………126

v. Ways to minimise the problems outlined above…………………128

11.4 Reasons for why SIML ratings were less accurate than SEQL ratings……130

11.5 Advantages of similarity ratings …………………………………131

i. Ratings provide a more nuanced picture of a witness’s memory…131

ii. Ratings provide three useful questions for juries………………132

iii. Ratings better represent reality…………………………………133

iv. The reduced import given to ratings may facilitate the collection of
    supporting evidence………………………………………………133

v. SEQL ratings provide a powerful new hybrid procedure for capturing
    eyewitness evidence that overcomes major problems associated with
    SEQL identifications………………………………………………134

iv. The use of an observation stage to benefit SEQL ratings and
    identifications ………………………………………………………137

11.6 Limitations of ratings………………………………………………138

i. Significant issue 1: Threshold responding and potential strategies for
    managing it…………………………………………………………138

ii. Significant issue 2: Targets were rated most similar less often than
    identified……………………………………………………………141

iii. It is unclear upon what basis similarity ratings are made…………141

iv. Ratings provide no conclusive statement of innocence…………141

v. The relationship between similarity and identity is imperfect……142

11.7 Future directions and applications…………………………………143

i. Applications…………………………………………………………145

Conclusion………………………………………………………………146

Cases cited………………………………………………………………149

References………………………………………………………………150

Appendices………………………………………………………………160
List of Appendices

Appendix A: Example Plain Language Statements........................................160
Appendix B: Graphic representation of Study 1 & Study 2 lineups.....................162
Appendix C: Graphic representation of Study 3 & Study 4 lineups.....................165
Appendix D: Questions for witnesses and jurors.............................................170
Appendix E: Example arrays............................................................................172
Appendix F: Example procedure for Study 1 similarity ratings (screenshots)...174
List of Tables and Figures

Figure 3.1: Factors affecting memory in the eyewitness context ........................................34
Table 5.1: Study 1 Experimental Design ........................................................................46
Table 5.2: Overall mean ratings (0-100) according to participant gender .................51
Table 7.1: Proportions of similarity ratings (study 1) and identifications (study 2) according to lineups 1-8 for target present and absent conditions ..........................................................60
Table 7.2: Example of a lineup demonstrating High-D and High-M on the left and Low-D and Low-M on the right. The proportion in which a lineup member was rated most similar (HR%) and/or identified is plotted on the Y-axis, with the lineup persons (1-8, or 9 = 'not present') plotted on the X-axis ........................................................................62
Table 7.3: Frequencies for levels of differentiation and match, and percentages for how well level of differentiation predicted level of match (Predictability%) ........63
Table 7.4: Example of extreme discrepancy (no relationship between similarity ratings and identifications) with High-D and Low-M. The person rated most similar (6) is not the most or second most identified .........................................................64
Table 8.1: Study 3 Experimental Design ........................................................................76
Table 8.2: Study 3 & 4 rating and identification data for targets/target replacements across visual/memory tasks and simultaneous and sequential procedures ........80
Table 8.3: T-test results for target present and target absent average ratings according to procedure and condition ..................................................................................84
Table 8.4: Frequencies and proportions for whether the target was rated most similar in the first compared to last lineup completed by participants .........................87
Table 8.5: The proportion of correct decisions based on the level of discrepancy between highest and second highest ratings for the memory task ..........................90
Table 8.6: The proportion of correct decisions based on the level of discrepancy between highest and second highest ratings for the visual task .................................91
Table 8.7: Discrepancy frequencies for target absent lineups for the memory and visual tasks and SEQL and SIML procedure ....................................................................92
Table 8.8: Paired samples t-tests comparing mean discrepancy scores in target present versus target absent lineups ..................................................................................92
Table 8.9: Univariate ANOVA results comparing mean discrepancies in target present lineups where the target was rated most similar versus a foil.................93
Table 9.1: Study 4 Experimental Design.................................................................95
Table 10.1: Frequencies for level of differentiation and match between the visual and memory similarity tasks and identifications..............................................103
Table 10.2: Frequencies for level of differentiation and match, and percentages for how well level of differentiation predicted level of match for the SIML and SEQL..104
Abstract

Eyewitness identification evidence is notoriously problematic, and the leading cause of wrongful convictions (The Innocence Project, 1992). The Australian Law Reform Commission has suggested that evidence of resemblance, rather than identity, is the most accurate evidence a witness can give. However, current practices in Australia continue to rely upon evidence of identity. This paper examines legal, philosophical and practical aspects of identity in relation to eyewitness evidence and proposes similarity as an alternative mechanism for gauging ‘sameness’. The four studies reported explored whether ratings of similarity offer an alternative method for capturing and presenting eyewitness evidence, and involved examining the relationship between similarity ratings and identification decisions in relation to unfamiliar faces. Study 1 (N=79) and study 3 (N=90) explored similarity ratings, while study 2 (N=67) and study 4 (N=57) examined traditional identification decisions. Results are reported for each study individually, but also compare study 1 with 2, and study 3 with 4. All participants completed eight photographic lineups online, which involved viewing a target face briefly, followed by a distractor face, and then viewing an eight-person lineup in which the target was either present or replaced by a similar looking target substitute. Participants then indicated whether the target was present or absent (identification condition), or rated the similarity of each lineup member to their memory of the target (ecphoric similarity), or rated the visual similarity between targets presented with the lineup (perceptual similarity). Between participant variables included the presence/absence of the target, lineup gender (male/female), lineup ethnicity (Caucasian/Asian), lineup procedure (simultaneous or sequential), task (memory or visual condition), and the position of the target/target replacement (early/late). Initial findings in regards to similarity ratings provide support for the hypothesis that similarity judgments underpin identification decisions. Advantages as well as some disadvantages of similarity are discussed. Further research is required, and recommendations are made accordingly.
Overview of thesis

Eyewitness identification is founded on the seemingly simple assertion that the person identified is the same person seen near or at a crime scene. Less clear is what underlies this claim of ‘sameness’. For an eyewitness to identify a suspect from a police lineup may appear a simple and intuitive task, but in actuality is highly demanding and involves complexities of memory, perception and identity. Recognition of this complexity, and that all was not well with eyewitness identification processes led to the formation of Devlin Inquiry in 1974 in the United Kingdom to “review in light of the wrongful convictions... all aspects of the law and procedure relating to evidence of identification in criminal cases; and to make recommendations.” (Devlin, 1976, Terms of Reference). The Devlin Inquiry was merely the latest in a series of such reports ordered by the Home Office on miscarriages of justice prompted by identification evidence (Shepherd, Ellis, & Davies, 1982). Similar concerns existed in the United States, where in 1967 the Supreme Court noted in United States v. Wade that “The vagaries of eyewitness identification are well-known; the annals of criminal law are rife with instances of mistaken identification.” The Devlin Inquiry determined that eyewitness identification evidence was inherently unreliable and that convictions should not be based on this alone. Significant recommendations for changes to pre and post trial eyewitness procedures were made and implemented. One such recommendation was that “Research should be directed to establishing ways in which the insights of psychology can be brought to bear on the conduct of identification parades and the practice of the courts” (Devlin, 1976, Section 8.1). Accordingly, the last 40 years have seen an explosion of psychological research into factors affecting eyewitness testimony (see Thomson, 2003 for a review). Despite the recommendations, eyewitness evidence has continued to result in the wrongful conviction of innocent persons. Sixteen years after the Inquiry, in 1992, the Innocence Project was established in the USA following a major study by the United States Department of Justice which found that incorrect identifications by eyewitnesses were a factor in over 70% of wrongful convictions (Innocence-Project, 2013). Through the use of DNA testing the Innocence Project has since exonerated over 300 people, 18 of whom spent time on death row. The extraordinary weight given to identifications in the criminal justice system is illustrated by the fact that faulty identifications were the sole factor leading to the jury’s decision in 50% of the cases. Furthermore, in 62% of these cases only one person identified the suspect as the perpetrator (Innocence-Project, 2013). Psychological research also suggests that identification accuracy
Identity Crisis in Identification Evidence

can vary wildly, with the false identification of innocent people fluctuating as much as 12% – 78% (Lindsay & Wells, 1985). Of further concern, identifications occur regularly: one United States survey from 1989 suggested that at least 80,000 eyewitness make identification of suspects in criminal investigations each year (Goldstein, Chance, & Schneller, 1989).

So why do wrongful convictions on the basis of mistaken eyewitness identification evidence persist despite almost 40 years passing since the Devlin Inquiry and the profusion of research into their causes? There are at least four reasons: first, research continues to indicate that eyewitness identifications are highly prone to error; second, despite this they are a powerful form of evidence that is afforded great weight by juries; third, despite the Devlin Inquiry’s encouragement for the courts to incorporate research findings, the “fundamental conceptual differences” (Melton, Petrila, Poythress, & Slobogin, 2007; p 4) that exist between law and psychology have slowed this process; and fourth, research involving identifications has largely accepted and adopted the flawed legal notion of identity and ‘sameness’ that underpins identification evidence. This paper will argue that this claim of sameness upon which identifications are founded is problematic, and that consequently, the current process for understanding and collecting identification evidence is unreliable. Despite its intuitive appeal, it is not helpful or even possible for a witness to make such a determination of identity, and the attempt to do so drives high wrongful conviction rates. Furthermore, when an innocent person is misidentified, the likelihood of apprehending the perpetrator is reduced. When a person is misidentified early in a case, valuable investigative resources may be misdirected. Thus, the illusion of precision surrounding identifications can cause problems both at the investigative and court stages (where identification evidence has great influence). Identification evidence can (mischievously) be compared to a used car salesman – full of superficial charm, promise and empty guarantees. Instead, similarity ratings as an alternative mechanism for gauging sameness are proposed. Rather than identifying one person from a lineup, this would entail a witness providing information on the visual similarity of all lineup members to their memory of the perpetrator.

There are two species of problem with identification evidence, one psychological the other legal. The psychological problem (described above) pertains to its being prone to error, which is due to a range of factors involving perception and memory. The second relates to legal process, and whilst significant, is typically overlooked. The legal problem is that with identification evidence the witness
assumes the responsibility of the jury. The juries’ role is to determine facts based on the evidence. For instance, to determine guilt of the accused jurors must establish three things beyond reasonable doubt: (1) that the offence occurred, (2) that it was intentional, or committed with a ‘guilty mind’, and (3) the accused is the offender. Current legal process allows the witness to potentially usurp the third role from the jury: e.g. the witness makes the determination of fact regarding the identity of the offender. However, it should be the jury that determines whether the identity of the offender is the accused. Thus, identification evidence is inconsistent with most other aspects of law. These psychological and legal problems have often been conflated, and the latter largely overlooked. The changes to the collection and presentation of eyewitness evidence proposed in this research aim to curtail both problems: they better represent reality and return the role of determining an offender’s identity to the jury.

This paper will integrate legal and psychological perspectives on eyewitness identification. In Chapter 1, identification evidence in the legal context is reviewed in terms of the Australian Uniform Evidence Act and relevant case law. Legal definitions, difficulties and a solution relating to eyewitness identifications are examined. Chapter 2 reviews the psychological research and current practice regarding identification procedures. Discussion covers the selection of foils and type of lineup presentation (simultaneous or sequential), as well as the two processes affecting identification decisions (discrimination and response criterion) and the factors that influence them. While some research pertaining to recognition of familiar faces is explored, the focus of this discussion and research relates to judgments regarding unfamiliar faces. Chapter 3 continues to examine the notion of identity and how it is determined from psychological and philosophical standpoints, and explores other factors affecting a witness’s ability to make an identification. Chapter 4 proposes similarity as an alternative to identity in eyewitness evidence, and a method for applying similarity to lineup procedures is introduced. The advantages and disadvantages of similarity are also discussed and the current research and hypotheses are described. Chapter 5 presents the first empirical study, which involves similarity. Chapter 6 presents the second study, which involves identifications. Chapter 7 compares the findings of these two studies. Chapter 8 presents the third study, which involves similarity, and includes several changes based on recommendations from the first study. Chapter 9 presents the fourth study, which involves identifications. Chapter 10 compares the findings of these two
studies. Finally, chapter 11 provides a general discussion of the research, its findings and applications.

CHAPTER 1.

Identification evidence in the legal context: Definitions and difficulties

Definition of identification evidence: The Uniform Evidence Acts

The law of evidence in Australia is a mixture of statute and common law. The Uniform Evidence Acts (Commonwealth Consolidated Acts 1995, and Victorian Current Acts 2008) represent an attempt to establish Uniform evidence legislation throughout Australia, though this process still continues today (Australian Law Reform Commission; ALRC, Report 102, 2006, s 2.1). As such, the Australian Uniform Evidence Act (UEA; Cth, 1995) is the statutory authority in relation to identification evidence. According to the UEA, identification evidence is evidence that identifies the defendant as being or resembling (visually, aurally or otherwise) someone who was at or near a place where an offence (or act connected to the offence) to be prosecuted was committed and must be based upon what the person making the identification saw, heard or otherwise perceived at that place and time (UEA Dictionary). The definition of identification evidence in the UEA does not include identification of a person other than the defendant, evidence about an object or item, and does not extend to civil proceedings. It also requires an ‘assertion by a person’, meaning that evidence of security surveillance footage or machine-based identification is excluded (ALRC, Report 102, 2006, s 13.9). The UEA makes an important distinction between two types of identification evidence, visual and picture identification evidence. Visual identification evidence is defined as “relating to an identification based wholly or partly on what a person saw...” (s 114.1). It is not admissible unless “an identification parade that included the accused was held before the identification was made” (s 114.2a), unless “it would not have been reasonable to have held such a parade” (s 114.2b), or “the accused refused to take part in such a parade” (s 114.2c). Picture identification evidence is defined as “evidence relating to an identification made wholly or partly by the person who made the identification examining pictures kept for the use of police officers” (s 115.1). It is not admissible “if the pictures examined suggest that they are pictures of persons in police custody.” Thus, the terms visual and picture identification evidence distinguish between whether a suspect was identified from a live identification parade (visual) or photographic lineup (picture).
Identity Crisis in Identification Evidence

Identification parades and photographic lineups

Given the inherent problems with identification evidence a number of mechanisms exist in law to ensure that identifications put before the court are as accurate as possible. Accordingly, the Evidence Acts “impose procedural requirements governing the way in which identification evidence is obtained prior to trial with a view to enhancing its reliability” (ALRC. 2006, s 13.6). The Devlin Inquiry indicated a preference for identification parades as the most reliable form of evidence. Under the UEA, the common law preference for identification parades becomes a requirement for admissibility of identification evidence. The two exceptions to this rule are outlined above (s 114.2b and s 114.2c). Picture identification is permitted in limited circumstances only and is subject to requirements, which seek to remove, or at least minimise, any unfairness to the accused (Alexander v. R, 1981; R v. Hallam and Karger, 1985; Evidence Act, 1995, Cth). However, in conflict with the UAE, a change made to the South Australian Evidence Act (1929) in 2014, states that identification evidence “is not inadmissible, and is not to be excluded, merely because it was obtained other than by means of an identity parade” (s 34AB). Also seemingly in conflict with the UEA, previous Victoria Police Manual instructions for the ‘Identification of suspects and offenders’ allowed for the use of photographic lineups in circumstances where “a police member merely has a suspicion about the identity of an offender” (‘Alternatives to identification parades’, s 5.1 – Photograph folders).¹ In the United States, while lineups are recognised as the most reliable form of evidence, they can involve either live persons or photographs, and there is no requirement that a particular form or procedure be used (ALRC, 102, 2006, footnote no. 72). In fact, in the United States, photo lineups appear to be more common than live lineups (Wogalter, Malpass, & McQuiston, 2004), with one survey indicating that 94.1% of responding law enforcement agencies reported using photo lineups. In comparison, 21.4% reported using live lineups (Police Executive Research Forum, 2013, p 48). Anecdotal evidence suggests that photographic lineups are also the norm in most Australian jurisdictions. In England and Wales standard practice involves the use of video or ‘viper’ parades (Valentine & Heaton, 1999). Viper parades entail nine short video clips where the witness views the front and sides of the head and shoulders of the

¹ A police member merely having suspicion about the identity of an offender as a reason for conducting a photographic lineup has been omitted from what appears to be an updated Victoria Police Manual. The relevant section in the updated manual is found in ‘Identifying offenders’, section 3 Verifying the identity of a suspect – Other methods (3.1 Picture identification).
accused and eight foils. Each person faces the camera at the start of the clip, then turns slowly to show their left profile followed by their right profile, before turning back to face the camera front on. The eight foils are selected from a database of 23,000 faces. The witness observes the video parade sequentially twice and then is given the opportunity of making an identification on a third run.

Photos are considered less reliable than live lineups for largely intuitive reasons; they provide less detail, exclude information regarding height, weight, movement, voice, and can sometimes be a poor likeness of a person. In the past they have often depicted the suspect in police custody, now recognised as incriminatory. An advantage of live lineups noted in Alexander v. R was that they involve the accused in the process, which is critical to the rights of the individual. For instance, when present in a live lineup, the accused (and their legal representative) can observe the process and any potential bias (e.g. that the police officer paused longer with the witness in front of the accused). Live lineups also allow for the accused to choose their position in the lineup, an aspect not afforded them with photographic lineups. However, given the time and cost efficiency of photo lineups they are regularly used in practice. There are also other advantages to using photos (Brewer, 2011). They do not require the witness and suspect to be in close proximity, making the process easier and less intimidating for the witness. The access to large photo databases makes it easier to compile an unbiased lineup, and these can be conducted on a computer, removing the need to have a lineup administrator present, which can introduce bias. Computers also allow for a standardisation of procedure, where witnesses are sure to receive the instruction that the perpetrator may or may not be present in the lineup. Use of computers also ensures a witness’s decision is recorded, whether regarding the suspect or another lineup member, and how long the witness took to make their decision. The research presented in this paper involves the use of photographic lineups, however, the findings and theoretical implications are envisaged to extend to live lineups.

The process by which a witness makes an identification is as follows:
Following a crime, when police have acquired a description from a witness or witnesses and compiled a profile of potential suspects, common practice is for an identification parade (or photographic lineup) to occur. This entails assembling one suspect and a number of innocents (foils), using either photos or real persons that match the witness’s description, or the suspect’s appearance. They are then viewed by the witness to determine whether an identification can be made, which involves making a binary (Yes/No) decision. There are two methods in which the lineup can
be presented to the witness; simultaneously, or sequentially. The simultaneous lineup (SIML) involves the witness viewing all lineup members (usually between 6-8) at the same time, and appears to be the current preference in most Australian jurisdictions (e.g. see the Victoria Police Manual, 2003; Australian Federal Police Practical Guide on Identification Evidence, 2007; South Australian Evidence Act, 1929; and New South Wales Police Force Procedures for the Evidence Act, 1997). In contrast, the sequential lineup (SEQL) involves the witness viewing lineup faces one at a time, and making a Yes/No decision for each face before viewing the next. There appears to be current debate in South Australia regarding the introduction of the SEQL procedure. In the United States, the emphasis has been on the reliability of the SEQL, rather than the SIML (National Institute of Justice, 1999), though a more recent report suggested that the issue of whether one procedure was superior remained ‘unresolved’ (National Academy of Sciences, 2014). Jurisdictional differences exist in regards to aspects of how the lineup procedure is implemented, e.g. in regards to the recommended number of persons/photos in a parade/lineup. Investigating police members should not take part in the procedure, however, seemingly in conflict with this the Victorian Police Manual notes that “if present at the parade, they should be in view of the suspect”. Most jurisdictions explicitly allow only one suspect in a lineup, though again Victorian Police guidelines indicate that “two suspects may be placed in the one parade if they are of similar appearance” though a “minimum of twelve other persons should participate in these parades”. Witnesses are not allowed to interact before or following the identification.

**Positive identification and resemblance evidence**

Other distinctions in terms of identification evidence have previously been proposed. In *Festa v. R* (2001), Justice McHugh discriminated between three forms of identification evidence; positive identification as direct evidence, positive identification as circumstantial evidence, and circumstantial identification evidence (Cavanagh, 2012). Justice McHugh described the first two forms as follows:

“A positive identification of the accused is direct evidence of the crime when it identifies the accused as the person who committed one or more of the acts that constitute the crime in question. A positive identification is circumstantial evidence when its acceptance provides the ground for an inference, alone or with other evidence, that the accused committed the crime in question. A witness gives direct evidence of the charge when she testifies that the accused ordered her to hand over the takings. A witness gives circumstantial evidence of the charge when she testifies
that the accused was the person who ran out of the bank immediately after other evidence proves it was robbed.” (Paragraph 54)

Justice McHugh described circumstantial evidence as follows:

“It is evidence that asserts that the general appearance or some characteristic or propensity of the accused is similar to that of the person who committed the crime. It may be evidence of age, race, stature, colour or voice or of a distinctive mark or gait. It differs from positive identification evidence in that the witness does not claim to recognise the accused as the person who committed the crime.. Although such evidence does not directly implicate the accused in the crime.. it is admissible evidence. It is proof of circumstance – usually, but not always, weak – that with other evidence may point to the accused as the person who committed the crime.” (Section 56)

Perhaps the most significant division between types of identification evidence entails the common law distinction between “evidence of resemblance”\(^2\), that a person shares certain features or attributes in common with the accused or looks or sounds like the accused, and “evidence of positive identification”, where the witness claims to recognise the defendant as the person previously seen (Festa v. R, 2001; Pitkin v. R, 1995; see also the Australian Law Reform Commission, Report 102, 2006). This division is apparent in Justice McHugh’s statements above.

In Festa v. R (2001) Justice Michael Kirby disagreed with Justice McHugh’s distinction between evidence of identification and resemblance, and suggested that the distinction was an artifact of the court. He stressed the importance of adhering to warnings regarding the potential dangers of identification evidence, arguing that such special protections were necessary for two reasons. The first concerned “the propensity of incorrect evidence of identity, even given honestly and with assurance, to involve mistakes leading to serious miscarriages of justice” (s 166). The second involved “the tendency for identification evidence to be given special weight, including in the mind of a jury.” (s 166). Such evidence links the accused to the crime, and no other evidence against the accused is then needed. Furthermore, creating a distinction between direct and circumstantial identification evidence “is likely to confuse judges, mislead lawyers and puzzle jurors” (s 167).

---

\(^2\) It is worth noting that the terms resemblance and similarity are used interchangeably in this paper and are assumed to refer to the same process.
Problems with the distinction between positive identification and resemblance evidence

There are several problems with the distinction between positive identification and resemblance evidence. These problems relate firstly to the consequences of the distinction in practice, and secondly, to the conception, collection and presentation of identification evidence.

Problems relating to consequences of the distinction between positive identification and resemblance evidence.

Two important consequences flow from the distinction between positive identification and resemblance evidence in practice. The first is that while evidence of resemblance alone is not sufficient to secure a conviction, instead forming part of circumstantial evidence (*Pitkin v. R*, 1995), positive identification alone is sufficient to ensure conviction; a matter over which the Devlin Inquiry expressed concern.

Given that faulty identifications are often the sole factor informing a jury’s decision, and more often than not include only one witness, this situation is problematic. As Justice Kirby noted, there is the possibility that positive identification evidence given by an honest and seemingly reliable witness can be incorrect, thus jeopardising the legal process. For instance, the link between confidence and accuracy is tenuous (Gardner, 1974), meaning that eyewitness evidence may be expressed in certain terms by a confident witness and still err.

Secondly, positive identification evidence triggers a warning from the judge to the jury regarding the potential dangers of eyewitness testimony (*Domican v. R*, 1992). This warning requires an explanation of the reasons for the need for caution, both generally and in regards to the circumstances of the case. It is not necessary that a particular form of words be used. Such warning can be critical to prevent miscarriages of justice, and to allow juries to make informed decisions. However, in *Festa v. R* (2001) the court suggested that warnings were not automatically required for evidence of resemblance. Subsequently, the situation can arise where evidence of resemblance is presented to a jury who receive no warning from the judge on the dangers of identification evidence. In practice, there has been some inconsistency in the application of whether and when a warning is given (see *Dhanhoa v. The Queen* [2003]; *Demiroz v. R* [2003], and *Trudgett v. R* [2008]) for cases involving dispute over judicial warning). This confusion reflects the attempt to find a compromise between the two extremes of making the warning mandatory or entirely discretionary. The recent Jury Directions Act (2015) clarifies directions in relation to
identification evidence in Victoria (see Part 4, Division 4, section 36). However, as noted by Justice Kirby, the distinction has created confusion in regards to when and whether a warning should be given. The result has been that in some cases a warning was not given.

Problems with the conception, collection and presentation of identification evidence.

The two consequences above flow out of what is potentially an arbitrary and unhelpful distinction between positive identification and resemblance evidence. However, perhaps a more fundamental problem relates to the assumption underlying positive identification evidence, which is evident in the way it is collected and presented. The assumption made regarding positive identification evidence is that it is possible to ‘recognise’ someone, with this recognition being based on a determination of ‘sameness’ between the person seen at a crime and in the lineup or parade. However, how this determination is made is unclear. Much of a person’s appearance can change dramatically, and even regarding identifications where a witness claims to know the suspect, they may be incorrect. 3 As the psychological research explored later in this paper will demonstrate, this assumption is neither simple nor straightforward, and there is much evidence to suggest that it is unhelpful at best, and unfounded at worst.

Another problem is that the process by which identification evidence is collected may not be representative of reality. It is not always clear whether an identification has been made. In the case of R v. Morgan (2009) "the question of identification... was pivotal in the trial" (s 10), which included confusion over whether or not the suspect had been positively identified. Accordingly, the jury needed to determine whether the identification evidence admitted was evidence of positive identification, or resemblance, and then gauge its reliability. There is a problem here in that the legal procedure may fail to reflect the witness’s experience. Whether completing an identification parade or photographic lineup, a witness is required to make a categorical decision (Yes/No) in relation to whether lineup members are the same persons seen previously. However, as in the case of R v. Morgan (2009), this imposition of a binary decision may not best represent the witness’s state of mind. The process of recognition likely occurs on a continuum, meaning forcing a binary

---

3 The distinction between recognition for familiar and unfamiliar faces is explored later in this paper (see page 30). Whilst it is possible for a witness to incorrectly recognise a suspect they believe is known to them, the focus of this discussion and research is on recognition of unfamiliar faces.
conclusion distorts reality and potentially produces misleading evidence. For instance, a tentative "maybe" may transform into 100% certainty with confirmatory feedback ("good, you picked the suspect"), and this can occur more often for inaccurate witnesses (Bradfield, Wells, & Olson, 2002). Thus, the combination of (1) potential perceived pressure by the witness to (2) make a forced binary decision that does not accord with their experience, may contribute to the high rate of false identifications and wrongful convictions. This is a radical point, as it suggests that the legal process surrounding the collection and presentation of eyewitness evidence is inherently flawed. Furthermore, the Yes/No decision limits the amount of probative information available to the court, thus undermining the ability of the jury to assess the evidence. Other methods can provide more probative information. For example, confidence ratings have been incorporated in eyewitness decisions in the USA, but have been discouraged in Great Britain and Australia due to being a poor predictor of accuracy (Leippe, 1980; Penrod & Cutler, 1995).

Melton, Petrila, Poythress, and Slobogin (2007) argued that there are fundamental differences between law and psychology with regards to the process by which each determine fact. They suggested that psychology typically employs probabilistic statements (e.g. 70% of people in situation X will behave in Y manner), whereas the law tends towards more absolute and categorical statements, such as guilty versus not guilty. They perhaps exaggerate the difference, for both psychology and law rely on probabilistic processes (in law this involves the standard of proof 'beyond reasonable doubt') to arrive at absolute outcomes (in psychology this involves whether a result was significant or not, hypotheses supported or not, or whether symptoms meet criteria for diagnosis or not). However, there are problems associated with reducing complex phenomena, whether they are psychological or legal, down to a twofold outcome. With identification evidence, the practice for the witness to be forced to make a binary decision regarding a suspect's identity likely contributes to faulty evidence in three ways. First, it perpetuates the illusion of precision and reliability in the minds of jurors that determinations of sameness upon which positive identifications are based are possible. Second, as noted, positive identification evidence is a powerful form of evidence that is afforded great weight in the minds of jurors. Furthermore, despite the potential inaccuracy of eyewitness testimony, a positive identification by one witness remains sufficient to obtain a conviction. Third, the witness usurps the role of jurors as determiners of fact. The role of jurors is to establish facts beyond reasonable doubt, and several facts are of particular importance in criminal cases. Namely, whether the offence being
prosecuted was committed, and whether it was committed by the accused. With positive identification evidence the witness is the one making the latter determination, meaning they are in a sense appropriating this role from the jury. As a consequence, identification evidence is inconsistent with most other aspects of law, and it undermines the role of the jury.

Evidence of resemblance is all a witness can give

Significantly, the Australian Law Reform Commission Evidence inquiry (ALRC: 2006 s 13.14) noted of a previous Evidence inquiry that "a suggestion was made that eyewitness evidence should only be permitted if expressed in terms of resemblance because a statement that the defendant "looks like" the perpetrator is the most accurate evidence that a witness can give. If a witness is under pressure to say "that's him", the witness may become convinced of the accuracy of the identification." However, this proposal was “ultimately rejected in recognition that it may weaken the force of sound identification evidence. There will be cases where the eyewitness can properly give more positive evidence, and such a limitation would prevent the witness from doing so." (ALRC, 2006, s 13.15).

Two things will be argued later in this paper in relation to this topic. First, that the suggestion made the ALRC that evidence of resemblance is the most accurate a witness can give is correct, and second, that the argument against it – that it weakens sound identification evidence – is misguided. Rather than requiring witnesses to make a binary decision regarding a lineup, which may represent a distortion of the underlying processes, it might be possible to utilise resemblance (or similarity) as a mechanism for gauging eyewitness evidence. Whilst the legal definition of identification evidence and the associated problems have been discussed, it is also necessary to examine the process from a psychological perspective. In the next section the current practices and relevant research in relation to the factors affecting eyewitness identifications are explored.

CHAPTER 2.

Current practice in identification procedures: Research and applications

In psychological research, factors affecting eyewitness identification have been separated into system and estimator variables (Wells, 1978). Estimator variables affect the accuracy of eyewitness judgments, but are beyond the control of the criminal justice system. They include factors surrounding the observed event,
such as the level of lighting, duration of observation and the distance between the perpetrator and witness. System variables are variables over which the criminal justice system can exert control. They include such factors as the lineup instructions given to witnesses, the likeness of foils (non-suspect lineup members; also termed fillers) to the suspect, the number of foils, the position in which the suspect is presented in the lineup, and the lineup procedure used (e.g. SIML or SEQL). Estimator variables are useful for providing information about the circumstances in which eyewitness evidence is more likely to be reliable. System variables are important for determining the most advantageous methods for collecting eyewitness evidence, and as such, are the focus here. Regarding system variables, Thomson (2003) noted that since the Devlin Inquiry, psychological research has resulted in two proposed changes to lineup procedures; the first relates to the way in which lineup members other than the suspect are selected (selection of foils), and the second to the method in which lineup members are presented to the witness, sequentially rather than simultaneously. While the first is described in part, the second issue is more the focus of this paper.

**The selection of foils**

There are two main methods for the selection of foils: match-by-description, where the foils are selected on the basis of the eyewitness’s description of the culprit; and match-by-appearance, where foils are selected by the lineup administrator on the basis of visual similarity to the suspect. The latter was employed in the research completed here. Match-to-appearance is the most common method (Wogalter et al., 2004), and whilst it was initially thought to protect innocent suspects from false identification (Lindsay & Wells, 1985), there has been suggestion that it actually increases this risk (Clark, 2003; Clark & Tunnicliff, 2001). One review of the relevant research (Fitzgerald, Price, & Oriet, 2013) reported that while the match-to-description procedure showed early promise, it showed little or no advantage over the match-to-appearance procedure in subsequent research. Furthermore, as outlined by Luus and Wells (1991), the match-to-description procedure is not viable when the description does not correspond with the appearance of the suspect, when the description is so specific that it is not possible to find foils who match it, or when multiple eyewitnesses to the same event provide contradicting descriptions. Additionally, witnesses can provide general descriptions, resulting in the situation where foils fit the description but look nothing like the suspect (Koehnken, Malpass, & Wogalter, 1996), or fail to provide any description of the suspects face whatsoever (Fitzgerald et al., 2013).
Simultaneous and sequential presentation

The second proposed change relates to the way in which the lineup is presented to the witness. As described above, this entails either simultaneous (SIML) or sequential (SEQL) presentation. Which procedure is more appropriate constitutes perhaps the most significant debate in the eyewitness identification field at present. The SIML involves the witness being given the opportunity of viewing all lineup members at the same time. This allows eyewitnesses to compare lineup members to each other and determine which most closely resembles, or is most similar to, the perpetrator - which is claimed by some to involve a relative judgment process (Wells, 1984). Under these conditions, Wells argued, there is a tendency for a witness to choose the lineup member who most approximates the offender, even when the offender is absent from the lineup. Having found that the SIML often resulted in the false identification of innocent people, Wells and associates developed the SEQL to discourage this tendency (Lindsay & Wells, 1985). The SEQL involves the witness viewing the lineup faces one at a time, and making a Yes/No decision for each one before viewing the next, with the lineup ending when an identification occurs (or all faces in the lineup being seen). Given that the eyewitness may think that the person being viewed looks more like the perpetrator than the last, they cannot be sure that next person will not look even more like the perpetrator. Thus, the SEQL was interpreted as reducing relative decision making (is this person more similar to the perpetrator than other lineup members?) and forcing a more absolute judgment process (is this the perpetrator or not?). Two basic experimental designs are possible using the SIML and SEQL procedures; the target may be present or absent. When the target is present three outcomes are possible: a correct identification (hit), the false identification of an innocent foil (false positive), and an incorrect rejection of the lineup (miss). When the target is absent, two outcomes are possible: a correct rejection of the lineup or an incorrect identification (either of a target replacement or foil). Lindsay and Wells (1985) found that whilst hits were equivalent across the SIML and SEQL, there was a significant reduction in false positives using the SEQL. They concluded that the SEQL could reduce false positives without significantly affecting hits.

A meta-analysis of 23 studies by Steblay, Dysart, Fulero, and Lindsay (2001) found that within the target present condition, the SIML resulted in more hits (50% vs. 35%), and less misses (26% vs. 46%) than the SEQL. In contrast, in the target absent condition, the SEQL resulted in more correct rejections (72% vs. 49%), and
Identity Crisis in Identification Evidence

less false positives (28% vs. 51%) than the SIML. Thus, the SIML has been considered more diagnostic of guilt when the target is present, and the SEQL more diagnostic of innocence when the target is absent. This is consistent with findings from police practice: the ‘Illinois study’ compared both the SIML and SEQL using actual cases from enforcement data and supported the finding that the SIML produces more suspect identifications and SEQL lineups, where as the SEQL lineup produces more non-identifications than SIML lineups (Mecklenburg, 2006; as cited in Malpass, 2006). An inconsistency is apparent between the results of Lindsay and Wells (1985) and Cutler and Penrod (1988), and those of Steblay and colleagues (2001; 2011). The former found the SEQL maintained hits whilst reducing false positives, yet the latter found that gains in reducing false positives were offset by reduced hits (participants were less likely to make an identification overall). It is unclear what underlies these contradictory results, and there is little comment on this discrepancy in the literature. The bulk of evidence certainly suggests an interaction between hits and false positives; one is improved at the others expense. Steblay et al. (2001) also noted that use of the SIML increased the likelihood of participants “choosing” (making any identification, whether correct or not).

Two factors affecting identification decisions: Discrimination and response criterion

Meissner, Tredoux, Parker, and Maclin (2005a) utilised a signal detection theory (SDT) framework to explore identification decisions. SDT examines the types of decisions that individuals make regarding old and new experiences, including their ability to correctly recognise an old experience (a hit response) and to falsely recognise a new experience (a false alarm response). By isolating these two decisions, SDT “separates an individual’s performance into two independent parameters – namely, discrimination accuracy (the ability of an individual to correctly detect a signal vs. correct reject its absence), and response criterion (the degree of evidence necessary for the individual to respond that a signal has been presented).” (p. 784). See also the report by the National Academy for Sciences (2014), chapter 5, for further discussion of this distinction. Thus with binary identification decisions these two processes affect witness responses and they are impacted by different factors. Factors that influence the quality of memory representation are thought to affect discrimination accuracy. This might relate to reduced capacity during the encoding phase (e.g. due to the own-race bias, own-gender bias, limited opportunity to view the perpetrator, or poor lighting) or storage phase (e.g. due to multiple similar experiences altering the memory). Alternatively, a person’s response
criterion is influenced by various social or instructional factors that may bias responding during the retrieval stage. The response criterion can be conceptualised as an internal threshold for choosing, and factors affecting it include lineup presentation (SIML or SEQL), similarity processes (e.g. the level of similarity between the target and foils), lineup instructions (e.g. that the suspect may or may not be in the lineup), the position in which the suspect is presented in the lineup (e.g. early vs. late presentation) motivation, familiarity and confidence, and external factors such as pressure to choose (Clark, 2005; Malpass & Devine, 1984; Meissner, Tredoux, Parker, & Maclin, 2005b; Steblay, 1997). Factors affecting discrimination are explored briefly. This is followed by a description of the factors affecting the response criterion, which is related to the current debate between the SIML and SEQL.

Factors affecting discrimination

Own-race bias.

One important variable affecting a person’s capacity to discriminate between faces pertains to the ethnicity of the witness and perpetrator. Own-race bias refers to the tendency in which people of another race than the eyewitness are harder to discriminate, and thus harder to identify accurately, compared to faces of the same race as the witness (Malpass & Kravitz, 1969; National Academy of Sciences, 2014). This effect is evident in research showing higher hit rates and lower false alarm rates for faces of a person’s own race compared to faces of another race (Meissner & Brigham, 2001). The combination of improved hits with reduced false positives implies an improved ability to discriminate for faces of one’s own race, rather than simply a change in response criterion – the latter would also result in increased false alarms. The inverse of the own-race effect (reduced capacity to discriminate for faces of other races) is evident in the fact that 42% of wrongful convictions based on mistaken identifications involved the witness making a judgment of a person of another race (Innocence Project, 2012). There is some evidence to suggest that reduced exposure time to the target increases the magnitude of the bias, expressed in terms of higher false alarms for faces of other races (Meissner & Brigham, 2001). Reduced exposure time would further diminish capacity to discriminate by limiting the amount of information encoded.

The National Academy of Sciences report into eyewitness identifications suggested that whilst the own-race bias is generally accepted, the reasons underlying it are less well understood. One explanation provided by Pezdek,
Blandon-Gitlin, and Moore (2003) is that repeated exposure to a stimulus allows an observer to accumulate a feature library pertaining to that type of stimulus. Thus, experience establishes a cognitive structure according to which new relevant information is encoded. This feature library or cognitive structure is used to match properties between objects. When fewer features are available for encoding, the encoded faces will incorporate less detail, and recognition will be more difficult. Thus, the ability to discriminate between faces may be reduced with ‘otherness’, and the proportion of features available in recognising a person from one’s own race is likely to be much greater than recognising a person from another race. The ethnicity of lineup members and participants were included as independent variables in studies 1 and 2 of this research. However, this was for the sake of comprehensiveness, and ethnicity did not constitute a major focus of this research.

**Own-gender bias.**

A similar, though less pronounced and consistent pattern exists in relation to viewing faces of one’s own gender. Palmer, Brewer, and Horry (2013) suggested that whilst females have consistently been found to be better at recognising female faces than male faces, the results for males vary. Some studies have indicated that males recognised male faces better than female faces (Ellis, Shepherd, & Bruce, 1973; Wright & Sladden, 2003), whilst others have suggested that males also better recognise female faces than male faces (McKelvie, Standing, St Jean, & Law, 1993; Rehnman & Herlitz, 2007). Gender (of lineup and participant) was included as a control variable in the research described below. Thus, similar to ethnicity, it did not constitute a major focus of this research. The own-gender and own-race biases also have implications for who selects the lineup using the match-to-appearance method. For instance, a male Caucasian police officer may generate a lineup of Asian female faces that differs substantially from one generated by a female Asian officer.

**Other factors affecting discrimination.**

Other factors that likely affect a person’s ability to discriminate include the duration for which a witness observes a perpetrator’s face during the commission of the crime (Bornstein, Deffenbacher, Penrod, & McGorty, 2012), the retention interval, or length of time between initial observation and identification (Dysart & Lindsay, 2006), and the level of stress and fear at the time of observation/encoding the perpetrator’s face (Deffenbacher, Bornstein, Penrod, & McGorty, 2004).
Factors affecting the response criterion

*Lineup procedure (SIML and SEQL).*

An alternative explanation to the relative/absolute distinction for the effect of lineup type on accuracy that draws on the response criterion was proposed by Ebbesen and Flowe (2002). They posited that the change in choosing rates (and differences in accuracy) brought about by the SEQL was caused by an alternation in participant’s response criterion. The response criterion is an internal threshold that must be exceeded for a witness to make an identification. The response criterion varies as a function of lineup presentation, but it also varies between individuals. Some people will naturally be more inclined to caution, others to impulsivity. An example of criterion change due to level of motivation might include witnesses being informed that false positives will result in a $50 fine versus being told that a correct identification will “help keep the community safe”. The witness will be less motivated, and consequently, less likely to make a selection, in the former situation than the latter. Meissner et al., (2005a) explored the impact of lineup type on response criterion, and found that across four experiments results consistently demonstrated a conservative shift in response criterion from the SIML to SEQL.

Often clouding research concerning SEQL is the absence of a consistent specification of the process. The wealth of research conducted means that a variety of methods and meanings have been applied. This causes the undesirable situation of a range of practices being implemented under the SEQL banner, as well as conflating comparisons between studies; it is impossible to compare apples and pears. In attempting to better understand the differences between SIML and SEQL, Zimmerman et al. (2006) broke the SEQL down into a “package” of variables, and identified two principal ones. The first variable was backloading, which occurs when witnesses are led to believe they will be presented with more photos than is actually the case. This is done to reduce the likelihood of witnesses making an identification if and when they realise the photos are running out; Lindsay, Lea, and Fulford (1991) found that knowledge of the number of lineup members in SEQL increased selections of an innocent suspect, leading them to recommend that witnesses not be aware of the number of faces to be presented. Zimmerman et al. (2006) suggest that the expectation that more photos will be presented can act as implicit instruction to hold off choosing, implicating a response criterion change rather than enhancement of witness memory.
The second variable was *multiple questions*, which refers to the number of identification decisions made by witnesses. In the traditional SIML there is only one question asked (“do you see the perpetrator in this lineup?”), whereas in the SEQL the witness is asked of every photo whether it is the perpetrator. As such the witness makes an identification decision about every lineup member seen in the SEQL, but only one overarching decision in the SIML. The relative/absolute explanation is that the one question in SIML facilitates an inter-photo relative comparison strategy, whereas multiple questions in the SEQL increase reliance on comparison between the photo viewed and the original memory of the perpetrator, thus explaining the reduction in false positives (Lindsay & Wells, 1985). Alternatively, the criterion change explanation proposes that the greater number of questions in the SEQL affects the responder’s response criterion. The reduction in choosing rates in SEQL compared to SIML suggests that multiple questions result in responders adopting a more conservative response criterion. However, it is possible that another variable is responsible for the reduction in choosing, and that multiple questions actually foster a more lenient criterion (e.g. encourage the witness to choose).

Zimmerman et al. (2006) found striking confirmation of the idea that specific package variables produce the differences between SIML and SEQL. When both SIML and SEQL had the traditional *simultaneous* package (non-backloaded, one identification question) no statistically significant differences were found. When both SIML and SEQL had the traditional *sequential* package (backloaded, 6 identification questions) no statistically significant differences were found. They concluded that the traditional SEQL package was responsible for higher correct rejections (and lower overall choosing) in sequential lineups. Removing the traditional SEQL variables eradicated the sequential superiority effect. However, adding the traditional sequential package to SIML did not cause significantly reduced choosing rates, and traditional and non-traditional SIML comparisons were not significantly different. Zimmerman et al. (2006) regarded this as an indication that the SIML was more robust to changes in procedure than the SEQL, an important consideration in real-world applications.

Following on from Zimmerman and colleagues, McQuiston-Surrett, Tredoux, and Malpass (2006) identified the *strict-stopping-rule* as another important variable in SEQL. The strict-stopping-rule ensures that once an identification is made the lineup procedure ends and no further faces are seen. If the strict-stopping-rule is not enforced then the witness continues to the end of the photos irrespective of an
identification being made, with the witness having the option of changing their choice. Findings are mixed as to the effect of the strict stopping rule. McQuiston-Surrett et al. (2006) reviewed 23 studies and found that a strict stopping rule decreased the SEQL ‘advantage’. In contrast, Lindsay et al. (1991) advised that witness’s should only view the lineup once after finding that a second sequential presentation did not lead to significant changes in identification decisions. Others have found that having two laps of the SEQL reduced overall accuracy compared to one lap (Steblay, Dietrich, Ryan, Raczenski, & James, 2011). Presumably, participant knowledge that the lineup ends upon identification will increase their response criterion as they implicitly realise they must be certain in order not to miss the real perpetrator. Yet strangely this was not the finding of McQuiston-Surrett et al.’s. (2006) review. However, only five of the 22 studies reviewed utilised the strict-stopping-rule, which may have affected results.

The relationship between discrimination and response criterion is blurred.

It is important to note that a number of the explanations above utilising the response criterion framework might also be explained in relation to discrimination. For instance, multiple questions might cue greater attentional focus leading to better ability to discriminate between faces. Inversely, factors attributed to discrimination, such as the own-race effect might involve the response criterion: if a person’s capacity to discriminate is reduced (all the faces look the same), then their preparedness to make an identification may also decrease. Thus, the distinction between discrimination and response criterion is not precise, and how they fit together is not always clear. This issue has not been clarified in the literature.

Similarity processes.

Similarity processes also affect the response criterion and discrimination. Flowe and Ebbeson (2007) predicted that foils low in similarity to the culprit lead witnesses to adopt a more liberal criterion for selection, which increases the likelihood that a suspect who looks similar to the culprit will be identified. They found that in both SIML and SEQL, a look-a-like was selected at a higher rate in lineups where the foils were less similar to a study face. The rate of choosing any face was also higher if the foils were lower in similarity to the study face, which the authors put down to a lower response criterion placement. One interpretation is that low similarity between lineup members results in a lower threshold for identification. Accordingly, foils dissimilar to the target may make identification of the target easier (improving discrimination), but may also reduce the threshold for choosing,
potentially increasing false identifications when the target is absent. Clark and Davey (2005) found that witnesses tended to select the foil most similar to the target when the target was absent, and that this occurred to the same degree for both lineup types. They also found that while there were no effects due to the positions of the target or next-best (NB) alternative in SIML, the ordering of the target and NB alternative played a large role in SEQL. More precisely, the target was identified often, and the NB rarely when the target was presented before the NB, and target and the NB were identified with equal probability when the NB preceded the target. Thus, the position of the target and other similar lineup members in the lineup impact upon the response criterion and identification rates. An alternative explanation involving discrimination is that when little difference in similarity is perceived between lineup members (e.g. they are highly similar), then discrimination between them is made more difficult. This would also explain Flowe and Ebbeson’s (2007) finding that a look-a-like was selected more often when foils were less similar to the study face.

**Lineup instructions.**

The instructions given to a witness regarding whether the suspect may or may not be present have been demonstrated to have a notable impact on a person’s response criterion. Malpass and Devine (1981) found that telling the witness that the suspect may not be in the lineup reduced the false positive rate from 78% to 33%. This highlights the importance of police instructions to the witness during the lineup process that the perpetrator may be absent from the lineup. However, whether such instructions go far enough is debatable, as a witness can validly presume that they have been called in for good reason. Somewhat of concern, a survey of police practices in the United States found that police volunteered using such instructions in only 50% of cases.

**The position of the target: order effects and counterbalancing.**

Another variable identified by McQuiston-Surrett et al. (2006) that differentiates between SIML and SEQL is counterbalancing; a way of safeguarding against the order of lineup presentation influencing the identification process. Counterbalancing involves presenting the target in multiple positions (e.g. early versus late) across the study design in order to reduce potential order effects impacting overall results. Previous studies have found order effects in lineups, demonstrating that counterbalancing is necessary. Clark and Davey (2005) noted a tendency in the SEQL (but not the SIML) for witnesses to pick the target more often
when presented early (position 2) compared to late (position 4) in the lineup. They argued that the explanation for this was straightforward; when the target was presented early, the NB alternative came after the target, and was less likely to be observed; when the target was presented late, the NB alternative was presented early and was more likely to be identified prior to the target being viewed (e.g. the identification was ‘spent’ prior to observing the target). This account utilises a response criterion explanation; the target and NB alternative exceed the threshold for choosing. However, it is likely that discriminability is also at play. For instance, the longer retention interval inherent to the SEQL procedure between viewing the target initially and subsequently in the lineup might also contribute to increased memory decay, reduced ability to discriminate, and poorer performance for identifying targets presented late. The greater number of faces viewed prior to the target when the target is presented late may also interfere with witness’s memory and impair capacity to discriminate. These differences are likely only relevant to laboratory studies, as in practice the greater duration (and number of faces viewed) between witnessing a crime and viewing a lineup would render this difference between the SIML and SEQL less meaningful. Similar to Clark and Davey (2005), Memon and Gabbert (2003) also found that participant’s choosing rates decreased when the target was presented late in the SEQL compared to the SIML - the target was exclusively presented in position 4 in their study. They also cited unpublished work by Ebbesen and Flowe from 2002 (not included in their reference list) that found that targets were more likely to be identified early in the SEQL. This was explained as being due to there being less opportunity for a foil to be selected without the target even being seen – e.g. one or two foils precede an early target compared to three or four preceding a late target, meaning there is more opportunity for a false positive for late presentation of a target.

In contrast to the above research, in a large study (N=2,529) of 24 comparisons between SIML and SEQL, Gronlund, Dailey, Goodsell, and Carlson (2009) found that presenting the target late (position 5) in the SEQL increased accuracy compared to early (position 2). They suggested this bias resulted from two things; first, witness reluctance to choose a face early in SEQL (higher response criterion at the beginning of the SEQL) due to being uncertain as to the level of variation between lineup members, and as to whether the real target was yet to appear; and second, increasing pressure on the witness to make a selection as the SEQL nears its end. One explanation offered was that participants were adjusting their response criterion after getting a sense of the variability between foils. For
instance, when the target appeared in position 2, the inherent difficulty and novelty of the SEQL could lead participants to initially have a high criterion. They also noted that two out of the three studies that found the largest SEQL ‘advantage’ in the literature placed the suspect last in an eight-person lineup (Lindsay et al., 1991; Lindsay et al., 1991). Carlson, Gronlund, and Clark (2008) also found that the SEQL advantage tended to occur when the target was presented late in the lineup. The analysis by McQuiston-Surret and colleagues found that the presence of counterbalancing decreased correct identifications for both SIML (from .59 to .44) and SEQL (from .56 to .25), and increased false positives for SEQL (from .24 to .39), but did not affect SIML. This indicates that the absence of counterbalancing increases correct decisions in SEQ (and to a much lesser extent SIML); a finding that suggests bias in lineup order is required to increase accuracy. It is unclear why the order of presentation in SIML would affect correct identifications given that all faces are viewed simultaneously, though the effect was much less pronounced than for the SEQL. Thus it appears that the SEQL is more susceptible to order effects than the SIML, and thus more reliant on counterbalancing in order to obtain an ‘advantage’. Order effects in the SEQ are likely to be affected by the similarity between foils and the target, and the position of the NB alternative. If foils are highly similar to the target, then identifications may be generally more likely to occur early in the SEQL, and thus when the target is presented early, indicate a bias for early presentation. However, if foils are less similar to the target, then the pressure to choose late in the lineup may bias late presentation of the target. The position of the NB alternative will interact with this bias. For instance, if foils are less similar to the target, and the NB alternative occurs prior to the target, then this may increase false positives of the NB alternative.

**Practical and theoretical issues with the SEQL**

The SEQL has increased in popularity and use over the past 20 years, however, there are suggestions that this change is premature (Gronlund, Andersen, & Perry, 2013). There are several reasons for this. First, certain package variables appear to produce the so-called SEQL ‘advantage’ (decreased false positives), and some of these package variables, such as the multiple questions, can be applied to the SIML. Secondly, the SIML may in fact be a more robust procedure; it is less sensitive to bias and the influence of extraneous variables. This is a worthy consideration, but one that is unheralded in the current debate between SIML and SEQL. Thirdly, the SEQL may be overly sensitive to order effects and in fact rely on a particular arrangement of the target and NB alternative in order to maximise its
Identity Crisis in Identification Evidence

‘advantage’. Thus it potentially involves inherent bias, and this has serious implications for practice; if law enforcement officers are aware that placing a suspect late in the SEQL will increase the likelihood of their being identified, then to position them late may prejudice the evidence, and to position them early may decrease the likelihood of a correct identification (assuming the suspect is the perpetrator). This issue is not resolved by simply allowing the suspect to choose their position. It is acknowledged that standard practice in England and Wales involves the SEQL viper (video) parade being viewed twice before and identification is made, meaning this issue is of less import in that jurisdiction. Fourthly, research involving the use of the signal detection framework has suggested that the simultaneous presentation of faces yields a higher discriminability than faces presented sequentially (Wixted & Mickes, 2014).

Quite apart from these practical issues with the SEQL, problems exist with the theory underlying it. The original rationale behind SEQL was that it prevented inter-photo comparison and forced comparison between the photo viewed and the witness’s memory of the perpetrator. This was intended to facilitate a more absolute and definite identification process. It may be on the basis of this separate comparison process that Lindsay and Well’s (1985) claim that SEQL could both improve hits and reduce false positives was made. Whilst intuitive, as well as popular in the literature, there are two problems with this account of the SEQL and the relative/absolute processes involved; it lacks evidentiary support and provides little actual explanation.

First, it is unsupported by the research: three meta-analyses conducted by Steblay et al. (2001); Steblay et al. (2011); and Gronlund et al. (2009) consistently indicate that eyewitness decisions are more accurate using the SIML when the target is present, and more accurate using the SEQL when the target is absent. When Carlson et al. (2008) replicated Lindsay and Well’s (1985) study, with the only difference being the selection of an innocent suspect that was less similar to the perpetrator, no SEQL advantage materialised. Gronlund et al. (2009) suggest that similar results may have been found regarding a SEQL ‘advantage’ because researchers conducted similar experiments. This lead Gronlund and colleagues to advocate SIML and SEQL comparisons across a wider range of variables. Furthermore, a report by the National Academy of Sciences in the United States proposed that the SEQL ‘advantage’ was partly a product of the statistical measure used (the diagnosticity ratio), which better satisfied the “popular criterion that those identified as guilty are actually guilty” (National Academy of Sciences, 2014). The
proponents of the SEQL have typically included only the selection of the target replacement in target absent lineups as a false positive (Lindsay & Wells, 1985). Thus, if the target replacement is presented in position 6, then the selection of foils 1-5 does not constitute a false positive. This was done on the basis that “the identification of a foil is a known error and does not function as a false identification in the true sense – for example, charges will not be brought against Officer Jones if he is identified” (p 557). Lindsay and Wells (1985) also argued that theoretically, false positives do not occur in target present lineups, as an identification of anyone other than the target is immediately recognised to be incorrect; what they term a ‘known error’. This exclusion of two types of false positives from the diagnosticity ratio measure (which involves dividing hits into false positives) can potentially create a misleading picture of the ‘accuracy’ of the SEQL procedure.

Second, the relative/absolute account actually offers little explanation; precisely how it explains the SEQL both maintaining hits whilst reducing false positives remains a mystery; how to confirm whether and when participants employ relative or absolute strategies is unclear; and it fails to take into account other factors known to influence hit/false positive rates. In comparison, the criterion shift explanation offers two advantages. The first is that it better accounts for the evidence. If the SEQL procedure results in a more conservative threshold for choosing (rather than an absolute judgment strategy) then any decrease in false identifications will necessarily be accompanied by a reduction in correct identifications, and this is generally the case (Steblay, Dietrich, et al., 2011; Steblay, Dysart, et al., 2011). The second is that it possesses greater explanatory power and is more easily tested. However, exactly how the relative/absolute and response criterion explanations fit together is unclear; the absolute/relative explanation for SEQL is either entirely separate to a response criterion change explanation, or it is a form of response criterion change explanation. Given that it was proposed of the original SEQL that it would both increase hits and reduce false positives in SEQL, it is likely that the relative/absolute theory is an entirely separate explanation. However, it appears to have gradually taken on the guise of a criterion shift explanation in the literature through references to participant’s increased "conservatism" in responding. Given that it is often couched in terms of increased conservatism, some may regard it as a form of response criterion shift. However, explaining reduced false positive rates using SEQL as a consequence of increased conservatism in an internal criterion is distinct from a relative/absolute explanation relying on separate comparison processes.
**SIML & SEQL: A misplaced debate**

The current debate between SIML and SEQL is misplaced on two fronts. First, instead of disagreeing which lineup is better, it may be helpful to focus on combining the strengths of each to create a robust and reliable hybrid. Many potential combinations exist. For instance, as the SIML appears more robust, a hybrid could consist of the SIML with traditional SEQL variables (e.g. multiple questions) applied. Secondly, and more importantly, while much, if not most of the research has focused on the differences between the SIML and SEQL, and examined the SEQL ‘advantage’, it has unquestioningly adopted the legal notion of identity (that forces witnesses into a binary decision process that may not reflect reality) and the assumptions underlying it. There has been little exploration of what underpins this notion of identity, and whether making such binary decisions are in fact possible. It is argued that a necessary shift must occur in the eyewitness evidence field, from accepting the current legal notion of identity, to applying psychological methods to better explore and understand it. The following section explores the nature of identity and the claim underlying identifications in more detail, as well as the myriad of associated problems. It then proposes similarity as an alternative mechanism by which eyewitness determinations may be made. Rather than making binary Yes/No judgments, this would entail witnesses rating every lineup member in terms of their similarity to the perpetrator. If, as suggested, the process of identification itself is unsound and responsible for high wrongful conviction rates, and other more basic similarity processes underlie eyewitness judgments, then the objective or rationale behind the SEQL is misguided. The issue is not how to best force an absolute identification process, but rather, how to better recognise, understand, and utilise similarity processes. Thus, in a sense the debate between proponents of SEQL and those less persuaded of its advantages is misguided. Instead of conforming lineup procedures to traditional legal concepts of identification, the traditional legal concepts themselves should be challenged.

As such, future research should focus on examining what underpins identifications, and exploring the role of similarity in lineup identification procedures. This will help determine whether identifications are the ‘used car salesmen’ of evidence – full of superficial charm and empty promises, but whose false guarantees should not be trusted. Incorporating similarity rather than identity as the construct underlying eyewitness judgments offers a number of significant advantages, which are explored in the next section. However, three immediate advantages that relate to the SIML and SEQL include the following. First, similarity
ratings could potentially reduce lineup order effects (e.g., the reluctance to choose a target early in a sequential lineup), and thus minimise the need for inherent bias in SEQL lineups. Second, similarity ratings may fulfill the function initially attributed to SEQL; they might reduce relative decision-making. A task requiring witnesses to “rate the similarity of each face viewed with your memory of the offender” would presumably draw less on perceptual similarity between lineup members (relative judgments), and more on ecphoric similarity comparisons between the lineup member being viewed and the witness’s memory of the suspect. Third, the requirement to rate similarity for every lineup member in a SIML would mimic one of the advantages of the traditional SEQL method; multiple questions.

At present, the author knows of no published research directly examining similarity as the mechanism underlying eyewitness judgments. Several studies have approached the issue. For instance, one study by Brewer, Weber, Wootton, and Lindsay (2012) that is explored later in this paper, whilst ostensibly exploring confidence, may have incidentally touched on the topic. Another study by Read, Vokey, and Hammersley (1990) found that an increased level of similarity between two photographs of a test person was associated with greater recognition accuracy. As such, there is a need for research exploring the role of similarity in identification procedures, and this should incorporate both the SIML and SEQL procedures. Again, the primary focus should be on challenging (or confirming) the notion of identity and identification currently employed by legal systems across the globe. Earlier in this paper, identification evidence was defined. However, the assumptions underlying identifications and the basis upon which the determination of sameness is made are not addressed in the law. These issues are now addressed in relation to psychological research. Problems with identifications are identified, and a more basic concept upon which identifications are likely based, similarity, is explored.

CHAPTER 3.
Identity in the psychological context: Definitions and difficulties

Defining identity and how it is determined

Identity is “the quality or condition of being identical in every detail; absolute sameness” (Oxford-Dictionary, 2002). The meaning in relation to eyewitness identification is clear: the proposition that the person being identified is the same (identical) person to that previously seen (at the crime scene). What is less clear is
how this determination is made. Given the shortcomings of face perception/recognition (that are explored in the following pages), this is an important question. A useful distinction in identity separates between an internal and external aspect. The internal aspect refers to the collation of internal experiences, associations and affiliations that constitute a person’s experience and sense-of-self independent of the physical body. The external aspect refers to exterior physical features or patterns, and importantly, is almost always the way by which the former is recognised or inferred. For instance, the face is the primary method by which friends recognise a person. Clearly other aspects are involved such as voice, touch, scent and movement, however, these too are external physical patterns that announce the inner person. Further complicating this process is the fact that the external aspect (face) by which we infer the internal aspect (person) is subject to change across time and space; a face can undergo great changes and remain the same person. The film ‘Face/Off’ depicts actor John Travolta undergoing a face transplant in order to infiltrate a terrorist camp. This is an extreme example of a common process; faces change over time. It also exemplifies the necessity of invoking an internal/external division where the latter can change independently of the former. Similarly, to look into the mirror 30 years apart is to see a different (external) personal looking back. A person is not their face, and the face is not the person, though we may unconsciously link the two in eyewitness identification procedures. Facial recognition, then, should be acknowledged to refer strictly to a collection of external features, and not a person.

The existence of an entity of “person” separate to or independent of the physical vehicle in which the person is housed has long been debated. Ancient Greek dualism was based upon a division between the body and soul. During the Enlightenment this demarcation was drawn between the rational mind and the physical body. In contrast, David Hume rejected the notion of personal identity existing over time, instead suggesting that the self is simply a bundle of (ephemeral) experiences. It is not necessary to invoke something immaterial to explain the internal aspect of identity; it can be conceptualised as an internal physical pattern (e.g. of neural networks relating to memory). To presume that only an external aspect of identity exists, is to suggest that (inevitable) facial/bodily changes over time correspond with changes in identity. However, the law clearly requires that identity be fixed. Otherwise holding an individual accountable for their crimes becomes impossible; an offender apprehended years after committing a crime might deny accountability due to being a different person to the one that offended. To
repeat, while it may be reasonable to associate a face with a person, it is strictly false to describe a face as the person.

This distinction between the two aspects of identity is necessary in eyewitness identification for the following reason. If a witness cannot identify the internal state of an unfamiliar person, then what is it they are identifying? Presumably they are recognising the external aspect – the physical configuration that constitutes a face (or body or voice). Yet, faces are commonly transformed (by time, expression, disguise, hair) meaning that relying on this external aspect in witness identification is perilous. This partly explains why identification evidence is often unreliable. To explore this in more depth it is necessary to examine how face identification occurs.

**Facial perception/ recognition**

Many researchers propose that faces are encoded via two systems (Bradshaw & Sherlock, 1982; Carey & Diamond, 1977; Chung & Thomson, 1995; Jensen, 1986); one system involves encoding isolated features, and the other the whole face. For instance, featural encoding (also termed part or analytic encoding) involves focusing on individual features such as the nose, lips or eyes. Alternatively, configural encoding (also termed global or holistic encoding) relies on encoding the relationship between features. There is evidence to suggest that people progress from feature-based to configural-based encoding with age (Carey & Diamond, 1977; Feinman & Entwisle, 1976; Mondloch, Le Grand, & Maurer, 2002; though a review by Chung and Thomson, 1995, suggested the opposite) and as a face becomes increasingly familiar (Ellis, Shepherd, & Davies, 1979; Veres-Injac & Persike, 2009). Thus, it appears that the eyewitness identification of an unfamiliar face relies on recognising particular external facial patterns that may be encoded using a piecemeal strategy relying on encoding features in isolation. Furthermore, this external physical pattern (face) is highly susceptible to distortion, both in terms of changes to the face itself, and the witness’s perception and memory of it. It appears that even reliance on the external aspect in identification is unsound.

For face encoding relying on a whole facial pattern, some researchers suggest a two-system model of face recognition that relies on separate perceptual (automatic) and cognitive (intentional) processes (Gillund & Shiffrin, 1984; Jacoby, 1991; Thomson, 1986). According to Thomson, faces are processed and recognised via two systems: “a hard-wired, automatic system which is perceptually driven on the basis of ‘feelings of familiarity’, and a cognitive system which brings to bear
analytic, synthetic and inferential processes.” These systems can work in tandem or separately; a person may initially be unable to recognise an old classmate’s face (familiarity), but be able to do so upon being told their name (cognitive/semantic).

A necessary caveat is that the focus of eyewitness identifications here will be on facial identity. Whilst other factors such as voice, movement, and body shape have been incorporated in eyewitness identification cases and been the subject of investigation, they remain beyond the scope of this discussion. It may be that the issues covered in relation to facial recognition apply to other methods of identification. If identification is based primarily on external pattern (face) recognition, then it is important to examine how robust memory for faces is.

Factors impacting on eyewitness identification

Bedford (2001) describes three challenges to making object identification decisions, which also apply to facial recognition; first, the movement of the eyes, head and body contribute to an ambiguous retinal image (e.g. a person is seen as small when distant and large when near); second, the objects of our perceptions themselves can move and change (e.g. when a person stands after being seated the image changes yet the pre and post-transformed stimuli refer to the same object), so when an object changes we need to both perceive those changes as well as know we are still dealing with the same object; and third, identifying enduring objects is further complicated by a lack of continuity of sensory stimulation. Stimuli go away and reappear when we blink our eyes, turn our attention elsewhere, or leave the room. Thus, we do not have continuity to tell us that a transformation, however extreme or ordinary, reflects the same object. These three challenges correlate with the main issues confronting face recognition. They are that the witness’s perception and subsequent memory of the offender’s face is influenced by multiple factors (distance from the offender, duration of viewing, lighting conditions), the facial features of the offender are subject to change (disguise, ageing, changed hair length or colour), and that these changes occur out of sight and cannot be determined by the witness.

Another method for conceptualising errors in face recognition employs the computer model of memory. This includes separating errors into whether they concern the circumstances at the time the offender was observed (the encoding stage), the intervening period between observing the offender and identifying them (the storage phase), or the identification methods and procedures (the retrieval stage). Factors relating to the encoding stage that have been implicated in impaired
identification accuracy include reduced lighting (Yarmey, 1986), brief duration of observation (Bruce, 1982; DiNardo & Rainey, 1991; Wells & Murray, 1983), whether there was a weapon involved (the weapon focus effect; see Steblay, 1992 for a review), the impact of stress on the witness (Valentine & Mesout, 2008), and the familiarity of the offender to the witness. Factors relevant to the storage phase involve the influence of intervening events on identification accuracy such as the impact on identification of prior descriptions (the verbal overshadowing effect; Schooler & Engstler-Schooler, 1990), and the impact on identification of exposure to “mug shots” (Cutler, Penrod, & Martens, 1987; Davies, Shepherd, & Ellis, 1979).

Factors affecting retrieval include the reinstatement of the context in which the witness observed the event (Smith & Vela, 1992; Thomson, Robertson, & Vogt, 1982; Wong & Read, 2009), such as returning to the scene of the crime, and method of lineup presentation (SIML or SEQL). See Figure 3.1 on page 34 for a visual illustration of the computer model of memory in relation to eyewitness identification evidence.

**The familiarity of the suspect to the witness**

One important factor in eyewitness identifications is the familiarity of the suspect to the witness. A great deal of evidence is mounting to suggest different mechanisms underlie the processing of familiar (famous or personally known) and unfamiliar (previously unencountered) faces. That familiar and unfamiliar faces might be processed in different ways was first suggested by Ellis, Shepherd and Davies (1979). While the focus of this discussion and research is on unfamiliar faces, it is useful to briefly review the literature for both. Johnston and Edmonds (2009) review evidence from neuropsychology, brain scanning, and psychophysics in suggesting that we process familiar and unfamiliar faces in both quantitatively and qualitatively different ways (see also Natu & O’Toole, 2011 for a review of the distinct neural underpinnings of familiar and unfamiliar faces). Recognition of unfamiliar faces is impaired whilst recognition of familiar faces remains largely unaffected by changes in viewpoint (Bruce, 1982; Hill, Schyns, & Akamatsu, 1997; Pourois, Schwartz, Seghier, Lazeyras, & Vuilleumier, 2005) expression (Bruce, 1982; Bruce, Valentine, & Baddeley, 1987), and context (Davies & Milne, 1982; Thomson et al., 1982; Watkins, Ho, & Tulving, 1976; Winograd & Rivers-Bulkeley, 1977). The definition of context varies between studies, though usually involves some or all of background, clothing, and orientation. For instance, Thomson et al. (1982, Experiment 5) found that recognition of unfamiliar persons was strongly affected by whether they were viewed with the same or different background,
clothing or orientation, but that changing these for familiar persons had no effect. They posited that in the familiar persons condition the fact that a known person was seen was encoded rather than information regarding the visual details of the face or context. One area of research has found that when identifying unfamiliar faces people rely more on peripheral features (e.g. the hair line, jaw line and ears) and alternatively, when identifying familiar faces focus more on internal features (e.g. the nose, eyes and mouth; Ellis et al., 1979; Hancock, Bruce, & Burton, 2000; Young, Hay, McWeeny, Flude, & Ellis, 1985). Furthermore, Bonner, Burton, and Bruce (2003) found that as faces became more familiar over a three-day period reliance on internal features increased. Eye-tracking studies also show that faces are processed differently depending on whether they are familiar or not (Thomson & Turnbull, 1989). There appears to be an unanswerable paradox here in that until a person is recognised it is unclear how it can be determined whether they are familiar or not. Consequently, it is uncertain how the eyes know where to look prior to this determination of familiarity.

The wealth and diversity of this research strongly suggests that facial recognition involving known persons relies on different information, is encoded differently, is more robust, and is more resistant to changes in context. Judgments regarding familiar faces are also more accurate (Roark, O’Toole, & Abdi, 2003; Veres-Injac & Persike, 2009), and this appears to be the case for humans as well as primates (Micheletta et al., 2015). As such, a witness identifying a person whose face is known to them would appear in many ways to be performing a separate task to a witness identifying someone not known to them. This raises questions over first, whether the two tasks should be given equal weight in the legal setting, and second, whether the same investigative processes should apply to both. For instance, one interpretation of the research described above is as follows: for cases in which the offender is known to the witness then the current process and terminology of “identification” may be more appropriate, but when the offender is unknown to the witness, then “identification” is too ambitious, and an alternative process is needed that draws on resemblance rather than identity. Should there be a separation between *identifications* – where the suspect is known to witness, and *resemblance judgments* – where the suspect is not known to the witness? This question cannot be answered with the current research, as it only pertains to judgments regarding unfamiliar faces. However, there are several reasons for discouraging separate eyewitness procedures for familiar and unfamiliar faces. First, a person may believe they have seen a person known to them, but be mistaken. Pezdek and Stolzenburg
(2014) found that judgments of familiarity were erroneous nearly 25% of the time. An earlier study by Young, Hay, and Ellis (1985) found that it was unusual for one familiar person to be misidentified as another familiar person (4.6% of all errors in facial recognition), but very common was the misidentification of an unfamiliar (not previously seen) person for a familiar person (86.7% of all misidentifications). What are being described here are perceptual errors occurring at the time of encoding (such as mistaking a stranger for someone familiar, or incorrectly encoding hair colour), rather than retrieval errors. The former cannot be remedied whereas the latter may be. Furthermore, in the situation where an incorrect identification is encoded there is little possibility for recognising the error. Further compounding this is the problem that subsequent storage and retrieval processes may be robust and the witness present as certain. A second reason to avoid separate identification processes for familiar and unfamiliar faces is that familiarity occurs on a continuum and is not categorical. Thus it is unclear where the line should be drawn and it would be arbitrary to do so. The case of R v. Morgan (2009) exemplified this conundrum. For instance, is familiarity determined by whether a person was seen previously to the witnessed event? Or by whether they were seen more than ‘X’ times previously? Or based on how long they were observed for? Or a subjective rating of familiarity completed by the witness? Third, such a distinction would require separate police and court procedures, placing a substantial burden on the criminal justice system. Fourth, a positive identification involving a familiar face would not assist in returning the responsibility of determining fact to the jury.

One way to incorporate judgments regarding familiar and unfamiliar faces into a procedure that maintains the juries’ role as fact finder would be for witnesses to provide an indication or rating as to whether they believe the suspect/perpetrator is known or familiar to them. Where the witness believes this to be the case, this fact could be presented to the jury, as well as the change in likelihood of the witness being correct in light of this information (judgments regarding familiar faces will likely be more accurate). Juries might also be advised of the mathematical possibility of the witness being erroneous in their judgment of familiarity (based on relevant research). Whilst more complex, such a process provides more information to juries, and accounts for all but the second problem outlined immediately above.

Given the internal and external aspects of identity, the impossibility of determining the former, and the myriad of factors influencing the latter, it appears that a conceptual shift needs to occur in the way facial recognition procedures are conceived and implemented in the legal context. “Identification” is too ambitious a
task, with the term creating an illusion of precision. The recommendation by the ALRC (2006) that judgments of resemblance are all that a witness can give appears persuasive. In contrast, its rejection on the basis that it would weaken sound identification evidence appears misguided. If identifications are flawed, unhelpful and even impossible, then an alternative is needed. One such alternative involves replacing the concept of binary identity decisions with continuum-based similarity judgments; where the witness rates the suspect (and other lineup members) in terms of their similarity to the offender. However, in order to do this, it is necessary to first determine the relationship between similarity and identity. As such, an important research question involves determining whether similarity judgments underlie identification decisions. If this is the case, then similarity and resemblance judgments may represent a more basic and useful measure by which evidence of resemblance can be collected and presented in court.

![Figure 3.1 - Factors affecting memory in the eyewitness context](image-url)
CHAPTER 4.

A solution: Similarity as an alternative to identity in eyewitness evidence

Introducing similarity

Similarity is of huge significance to psychology, and there is a vast and complex literature dedicated to it. Medin, Goldstone, and Gentner (1993) suggest that similarity is a comparative process that is based on matching or mismatching properties, and that two things are similar to the extent that they share properties, and dissimilar to the extent that properties apply to one entity and not the other.

Past research on face similarity has typically relied on a geometric model of similarity where the components of a face are treated as points in a multidimensional space. The similarity between two faces is determined by the distance between those two points (Ross, 2008; see also Valentine, 2001, for a review). In attempting to conceptualise and implement a measure of facial similarity based on the physical properties of the face, Tredoux (2002) employed a computer based geometric model called Principal Components Analysis (PCA). The study examined the relationship between similarity measures derived from PCA and human estimations of face similarity. Results suggested that PCA corresponded reasonably well with human judgments of facial similarity. Tredoux's (2002) aim was to operationalise similarity and make it precise and testable, thereby providing a framework for future research. He concluded that PCA had potential application as a software tool for constructing arrays of faces of varying similarity as well as for reconstructing facial images from memory. A limitation of precise models such as PCA, and one acknowledged by Tredoux, is that they fail to account for the subjective element of similarity. Perceptions of what constitutes similarity vary between individuals. For instance, opinions of whether a child resembles their mother or father differ. Thus, a fundamental aspect of similarity is that it is relative.

While Tredoux focused on generating similarity, another focus might be to utilise witness’s perceptions of similarity as a criterion for guiding and describing eyewitness decisions. The latter application would better account for the subjective nature of similarity.

Similarity and Lineups

Tredoux examined similarity comparisons between two faces that were presented simultaneously, however, other similarity comparisons occur. In several
studies of recognition accuracy using complex pictures, Tulving (1981), distinguished between two types of similarity, perceptual and ephoric. Perceptual similarity refers to the similarity between test items in a set (both are physically present). Applied to lineup identification procedures, perceptual similarity is how similar the lineup members are with each other (akin to Tredoux). Ephoric similarity on the other hand refers to the similarity between a test item and the stored relevant episodic information (one is physically present, the other exists in memory). In lineup procedure terms, ephoric similarity refers to how similar lineup members are with characteristics of the perpetrator in the witness’s memory. Ephoric similarity can occur when a lineup member is similar to another person at the scene of the crime, or a person seen previously. This means that eyewitness decisions can err in relation to ephoric similarity (a witness mistakenly identifies a foil who resembles the offender), and/or perceptual similarity (the influence or bias of other lineup members on the eyewitness decision). Given that ephoric similarity occurs between a lineup member and the witness’s experience (something not physically present), it is difficult to measure. Perceptual similarity, however, occurs between two things that are present (lineup members) and can be controlled. The court runs a risk here in relation to witness identification. It is that an identification is not made on the basis of ephoric similarity, but on the basis of perceptual dissimilarity; a suspect is identified not because of their similarity to the offender, but as a function of their dissimilarity to other lineup members – they stand out in the lineup. An anecdotal example entails an identified suspect being the only lineup member to have a tearstained face. Conversely, an innocent person may be identified for approximating a witness’s memory of the offender relative to other lineup members. Thus, the witness’s perceived similarity between the offender and the suspect is relative to, or dependent on, the similarity of the members of the lineup. Presumably the court requires that there should be a minimum standard of similarity, of which less (e.g. that both the offender and suspect have blue eyes) is insufficient. But it is difficult to exclude the possibility that the basis for similarity is a function of the lineup.

Research on similarity in relation to lineups has focused on perceptual similarity, and two issues in particular; lineup fairness (or lack of bias) and accuracy (increased hits and reduced false positives), and there may be a trade-off between the two. As suspect-foil similarity increases lineup bias decreases but accuracy can also decrease (Malpass & Devine, 1983; Tredoux, 2002). Thus, as lineup members are increasingly similar to the suspect the lineup becomes fairer, but identifying the
suspect more difficult. However, findings regarding the effect of suspect-foil similarity on accuracy are mixed, see Ross (2008) for discussion.

Common to most accounts of similarity (Medin et al., 1993; Tversky, 1977) is the idea that inherent to judgments of similarity are judgments regarding dissimilarity. This is not to suggest that similarity and dissimilarity judgments necessarily represent the inverse of each other, as there is evidence to suggest that they may not (Medin, Goldstone, & Gentner, 1990). For objects of greater likeness dissimilarity is a more relevant distinguishing feature, whereas for unlike objects similarity becomes more pertinent in separating between them. In his work incorporating non-facial stimuli, Osgood (1949) described the paradox of similarity; the idea that the ability to discriminate between two objects increases in difficulty as they become more similar (resulting in more false positives), until they become the same (at which point it becomes a hit). Identical twins exemplify this problem in the eyewitness context. Similarity increases to the point of a hit (in terms of similarity), yet remains a false positive (in terms of identity). One interpretation of Osgood’s paradox is that similarity processes underlie identification judgments; that decisions of identity are made on the basis of similarity. Thus, it may be that similarity is not only a more useful and probative tool in guiding and describing eyewitness decisions, but also the process upon which identifications are inherently made.

If increasing the similarity between two faces increases the difficulty in discriminating between them, so too will reducing the dissimilarity between two faces make discrimination more difficult. For instance, faces for which fewer features are discernable will be more difficult to distinguish. This might be due to external factors such as poor lighting, or cognitive factors. As noted earlier, Pezdek and colleagues suggested that repeated exposure to a stimulus allows an observer to accumulate a feature library pertaining to that type of stimulus (Pezdek et al., 2003). This feature library or cognitive structure is used to match properties between objects (make a similarity comparison). When fewer features are available for encoding, the encoded face will incorporate less detail, and recall/recognition will be more difficult. This reinforces the subjective nature of similarity. If one does not have the conceptual library that facilitates a similarity judgment, then one is less likely to perceive similarity between two objects.

In a study exploring the effects of exposure duration and photo similarity on a recognition task not involving lineups, Reid, Vokey, and Hammersley (1990) separated two target photos into low, medium or high similarity. The two photos in the high similarity group were identical. In a subsequent recognition task participants
were instructed to respond ‘Yes’ or ‘No’ for each slide presented. Results indicated that when similarity between the two target photos was high, increased exposure duration always improved subsequent recognition. When the two photos were of low or medium similarity to each other increased study time had either no effect, increased performance, or in some cases significantly reduced performance. The authors interpreted results as suggesting that as exposure time increased, participants encoded more view-specific than face-specific information, and that reliance upon this information only aided recognition when the similarity between the two photos was high. Another interpretation is that increased ecphoric similarity (occurring in the high similarity condition) resulted in greater recognition accuracy.

**Making similarity operational**

So returning to the question of how similarity relates to identity, it might constitute a threshold, where a certain quantity of similarity (summed) triggers an identification. This is conceptually very different to a witness “recognising” a person. Put another way, identification decisions may represent an arm wrestle between similarity and dissimilarity thresholds. In a study tracking the eye movements of participants, Flowe and Cottrell (2011) distinguished between patterns suggesting either an automatic recognition process or a more deliberate comparison strategy. They suggested there may actually be two thresholds for choosing; one for a positive identification and another for a rejection, and when a face falls between the two thresholds, deliberation occurs. If identification processes do mimic similarity processes, then (even putting all the difficulties associated with identity aside) it is misleading to force eyewitness’s decisions made on a continuum of similarity into a dichotomous identity decision of Yes/No. A more appropriate method would allow for keeping similarity judgments on a continuum such as a rating scale (e.g. ranging from 0 to 100% similarity). A percentage based-scale may present a more intuitive measure that renders the task easier and thus better facilitates witnesses’ judgments. Such a scale is incorporated in study 1 below. Examining this proposal involves investigating whether similarity and identity judgments follow a similar pattern in eyewitness identifications using lineup procedures. The first step in determining this constituted one of the main research questions in this study. The question explored was: “Is the person identified most often also rated as being the most similar?

Brewer, Weber, Wootton, and Lindsay (2012) conducted a study where participants were asked to provide quick (within three seconds) confidence assessments (0-100%) for each face presented in a sequential lineup. Confidence
ratings pertained to whether a face had been presented earlier (e.g. in a study phase). Results suggested that the deadline confidence ratings procedure produced significantly higher classification accuracy than did a control condition using the traditional Yes/No identification procedure. Furthermore, large differences between the maximum and next-highest confidence ratings (e.g. 70% - 100%) denoted very high accuracy; in contrast, small differences (< 20%) denoted low accuracy. Thus, an individual's confidence profile provided useful probative information that was not possible using the binary decision procedure. One explanation is that requiring participants to quickly rate confidence rather than make an identification forced them to rely more on global than local processes. Another study by Macrae and Lewis (2002) used stimuli consisting of large sized letters (e.g. H) made up of smaller sized letters that were either the same (H) or different (T). Participants were primed towards global processing by being asked to respond to the large letter. They were primed towards local processing by being asked to respond to the small letters. In a subsequent SIML task in which controls yielded 60% correct identifications, the global processing condition performed at 83%, while performance in the local processing condition dropped to 30%. Similar results were found by Perfect, Dennis, and Snell (2007).

The confidence ratings made by participants in the Brewer et al. (2012) study were described as “rat[ing] the degree of match between the culprit and each lineup member” (p. 1209). Given that these participants did not make an identification (accuracy being related to the highest confidence rating), these confidence ratings may have in effect been similarity ratings of sorts; a rating on a 0-100 scale of confidence that a face has been seen previously may be an indirect measure of similarity. There may be an important distinction between a post-identification rating of a witness's confidence in their decision (as commonly used), and the confidence ratings used by Brewer and colleagues, in which the identification was replaced with a rating of the witness's confidence that they had seen the person previously: the latter more likely approximates a similarity rating. In fact, other studies by this research group have even linked their confidence ratings with ecphoric similarity (Sauer, Brewer, & Weber, 2008; Sauer, Weber, & Brewer, 2012), and suggested that confidence ratings may provide a "relatively direct measure" of ecphoric similarity (Sauer et al., 2012, p. 490). It is unclear why these studies have incorporated confidence and not similarity as the mechanism upon which to ask participants to base ratings, particularly given the references to ecphoric processes. This may in part be due to confidence ratings already forming
part of the literature. However, it would appear that inferring similarity from confidence ratings results in a less direct measure.

Providing ratings rather than binary decisions may not only provide more nuanced eyewitness decision information, but may also change the way in which participant’s respond, for instance by removing the pressure of making an identification. The problem of false positives is in one sense eliminated (as no person is identified), yet probative information regarding the lineup member most similar to the perpetrator is retained. Comparisons between ratings is also possible; if ‘person 3’ corresponded to a confidence rating of 84%, and the next highest was ‘person 4’ on 55%, then this is more revealing than if the next highest rating is 78%. Discrepancy between ratings provides information on the witness’s spread of decisions and can index more probative information. Thus, the similarity rating given to a suspect is interpreted in the context of those given to the foils, thereby providing a more comprehensive picture. At minimum, Brewer et al.’s (2012) study suggests that measures other than identification may be used effectively in the eyewitness context. Remarkably, the very process of requiring an identification was demonstrated to change the way participant’s responded. One explanation for this is that identification is a secondary process that is artificially superimposed over other processes (e.g. familiarity, similarity).

It should be noted that it is unclear what the impact of making time-pressured decisions was on the Brewer et al.’s (2012) results. Should participants have been given more time it is possible that they may have responded differently. Additionally, in Brewer et al.’s (2012) study participant’s confidence ratings were based upon sequential lineups only. Consequently, future research should explore Brewer and colleagues’ basic paradigm in regards to three key elements; first, using similarity ratings instead of confidence ratings, second, eliminating the time-pressure aspect to determine whether similar results occur, and third, incorporating simultaneous lineups. The research explored below incorporated these three features.

**Advantages and disadvantages of similarity**

There are numerous advantages for replacing binary identification decisions with similarity ratings as the mechanism upon which eyewitness decisions are based. First, as mentioned, ratings allow for a more nuanced description based on a continuum (for instance a rating scale of similarity) that provides more information than a mere Yes/No identity decision. Second, it removes the burden or
responsibility on the witness to make an artificial dichotomous decision that they
may be unable or uncomfortable to do. Citizens naturally desire to assist the court,
and it may be that forcing an identity judgment on what is a judgment of similarity
contributes to the wrongful conviction of innocent people. For instance,
identifications made when the witness’s response criterion is operating near
threshold may be more prone to error, however, the Yes/No response provides no
information on this. Using similarity ratings, the issue of threshold responding is
better recognised, even if not removed. Brewer et al. (2012) suggest that the
perceived significance of the identification decision (as opposed to rating) may
change the way participant's respond. A related point is that similarity ratings reduce
the traditional bane of eyewitness evidence – false positives. No person is identified
using similarity ratings, yet probative information regarding similarity of and between
lineup members is retained. Third, ratings assist in returning the role of fact finding
to the jury where it belongs, rather than forcing it upon the witness. This is
important, as using similarity ratings will not eliminate the potential for innocent
persons to be rated as more similar than other lineup members, but it will eliminate
the aura of certainty surrounding a positive identification in such cases, and
importantly, return responsibility to the jury to determine the reliability of the
eyewitness evidence. Fourth, it would help reduce the illusion of precision currently
surrounding eyewitness identifications in the public mind. For example, changing the
requirement for a witness to rate similarity, rather than make an identity judgment,
as well as changing the terminology from a “person identification” to an “eyewitness
decision” task, would help make salient to those involved (particularly the jury) the
risks and fallibility of eyewitness judgments.

One disadvantage of using similarity constitutes an advantage inversed.
Describing identification evidence on a continuum rather than category renders it
probabilistic rather than an absolute. Despite being less representative of reality, the
binary identification process is by its nature simpler and easier to interpret; the
illusion of precision is a comfortable one. As such, simplifying the process might be
argued to reduce the burden on juries. There exists here a tension between the
ease of process and the integrity of process. A second limitation of similarity is that it
is theoretically possible for an identification to occur without reference to similarity.
In one particular case the identity of the defendant was determined by inference: a
young woman was seen entering but not exiting a public toilet, and an old woman
was seen exiting but not entering (D. Thomson, personal communication, March 6,
2015). Thus, a criticism of using the similarity ratings procedure advocated here
might be that it does not apply to all situations. However, the above example is highly unusual, and the similarity ratings procedure can be applied to eyewitness evidence involving the use of lineups or parades (which is the vast majority): it is comparable to the “evidence of resemblance” described in *Festa v. R*, 2001 – that a person looks or sounds like the accused. Alternatively, a person can be perceptually dissimilar and still be the same person. Thus there may be some (rare) instances where visual similarity ratings do not underpin identification decisions. It appears likely that similarity becomes more important, and is the likely basis for identifications, as a face is less familiar (eradicating the possibility of an identification based on feelings of familiarity). This research examines the relationship between similarity and identity for unfamiliar faces, thus it can be expected that a stronger relationship will exist than if familiar faces had been used.

A final limitation of similarity ratings is that they do not necessarily answer the question of how similarity was determined.

The current research and hypotheses

The general aim of this research was to explore the possibility for similarity ratings to provide an alternative to identifications as the method for obtaining and presenting eyewitness evidence. As yet there are no known published studies directly exploring similarity ratings as an alternative to identifications, thus most of this research was exploratory in nature. As such, the research involved fewer predictions and more questions. In sum, the four studies reported here addressed the following research questions. (1) Are similarity ratings related to identification patterns? The first step in answering this involved determining whether the person rated most similar was also identified most often in the corresponding identification condition. (2) Do similarity ratings provide a useful alternative? In order to offer an alternative to identifications, it is necessary for similarity ratings to be diagnostic of accuracy, and therefore useful. This entailed asking the following question: Are targets rated more similar than target replacements and foils? (3) Also of interest was asking: Under what conditions are ratings most diagnostic of accuracy? For instance, it was expected that greater discrepancies between the highest and second highest similarity ratings would index increased accuracy (the target being rated most similar). In contrast, it was expected that low discrepancies between the highest and second highest ratings would index lower accuracy. (4) Finally, given the exploratory nature of the research, an important question was: How do similarity ratings differ across the experimental conditions (target present/absent, early/late
presentation, SIML/SEQL presentation, male/female lineups, Caucasian/Asian lineups)?

Four experiments were completed overall, with the following aims and rationale. Study 1 provided an initial investigation of similarity ratings with regards to the SIML procedure only. The focus was to explore the nature of similarity ratings and whether the target was rated most similar. Participants in study 1 completed similarity ratings for eight SIML lineups of male and female, Caucasian and Asian faces. The presence/absence and order of presentation (early vs. late) of the target was varied. Study 2 replicated study 1, but participants made traditional identification decisions rather than similarity ratings. The purpose of study 2 was to provide a basis for comparison with study 1. Of interest was determining whether the person rated most similar in study 1 was identified most often in study 2. Also of interest was determining whether more targets were correctly identified than rated most similar (comparative rates of accuracy with identifications and ratings). Thus, studies 1 and 2 formed part of a larger research project. Study 3 involved similarity ratings, though with an expanded focus and including recommendations from study 1. The aim was to continue to explore ratings in relation to additional conditions, as well as determine whether the results of study 1 would be replicated. Participants in study 3 completed similarity ratings for eight lineups of male Caucasian faces. Both the SIML and SEQL procedures were included. A new ratings condition based on visual (perceptual) similarity was included in addition to the memory-based ratings employed in study 1. The presence/absence and order of the target were again varied. Study 4 replicated study 3, but participants made traditional identification decisions rather than similarity ratings. The purpose of study 4 was again to provide a basis for comparison with study 3. Thus, studies 3 and 4 also formed part of a larger research project. The same questions of interest explored with regards to comparing studies 1 and 2 were addressed with regards to comparing studies 3 and 4. Results are reported separately for the four studies, but further results sections are included that compare study 1 similarity ratings data with study 2 identification data, and also study 3 similarity ratings data with study 4 identification data. However, discussion sections for each individual study are not provided. Instead discussion sections are provided after the comparison between studies 1 and 2, and studies 3 and 4.
CHAPTER 5.
Study 1
Method

Participants

Two-hundred-and-six persons accessed the survey page, of which only 38% completed the study. The vast majority (77%) dropped out in the practice lineup stage. Data was cleaned, with non-completers excluded from analyses. Also excluded were any participants who provided zero ‘not at all similar’ responses across all lineups. There are several reasons for the high rate of non-completers. High dropout rates are a well recognised limitation of online research (Dandurand, Shultz, & Onishi, 2008; Frick, Bachtiger, & Reips, 2001); in laboratory settings participants feel more compelled to remain and finish an experiment, whereas in online studies they can leave at any time. Furthermore, the ratings task was quite taxing (e.g. it required eight responses per lineup), which is indicated by the high dropout rate in the practice lineup stage. Other than participant drop out, missing data were likely due to technical problems pertaining to Internet connections, as the program required that a person enter all requisite responses prior to moving on to the next task.

The final sample therefore comprised a total of 79 participants who completed study 1. They were aged 16-58 years ($M = 26.15, SD = 9.1$), and 74.7% were female. Caucasian participants constituted 55.1% of the sample, with 27.8% being Asian, and 17.7% of another race. Approximately 18% were born in Australia, with the remainder born overseas. Participants were recruited through email invitation, social media sites (e.g. Facebook), online psychological research sites, and through word of mouth.

Design

Study 1 entailed memory-based similarity ratings involving the SIML procedure. Within participant independent variables included the target presence/absence, lineup ethnicity and lineup gender. Between participant independent variables included early/late presentation, participant ethnicity and participant gender. Thus, study 1 utilised a 2 (target: present/absent) x 2 (lineup ethnicity: Caucasian/Asian) x 2 (lineup gender: male/female) x 2 (order of target/target replacement presentation: early/late) x 2 (participant ethnicity: Caucasian/Asian) x 2 (participant gender: male/female) mixed design. In the target absent condition a target replacement (determined by the experimenter to be
visually similar to the target) was substituted for the target in the lineup. Early presentation entailed the target/target replacement being presented in the second or third position of the eight-person lineup. Late presentation entailed the target/target replacement being presented in the sixth or seventh position. A different photo of the target was presented initially and at the lineup stage. The two photos of the target were taken on separate days and differed in terms of background and clothing.

The dependent variable constituted participants’ ratings of the similarity between each lineup member and the participant’s memory of the target. Ratings were made on a 0-100% scale, where 0 represented “not at all similar” and 100 represented “extremely similar”. The order of lineups was randomised and all participants completed four target present and absent, early and late, male and female, and Asian and Caucasian lineups (see Table 5.1). The order of all foils but one was held constant. One foil was swapped with the target/target replacement in the early and late conditions. For instance, if the target was presented early in position 2, the foil appeared in position 6, whereas if the target was presented late in position 6, the foil appeared in position 2. Two lineup compositions were included in the study (see Table 5.1). If a participant completed the first four lineups in which the target was presented early for each race and gender, then in the latter four lineups the target would be absent and a target replacement presented late. This meant that for any one participant the order variable was not fully crossed. For example, in condition 1, a participant completed target present/early lineups (1-4), and target absent/late lineups (5-8), but did not complete target present/late, or target absent/early lineups. Participants in condition 2 completed the latter two.

It was not possible to perform one analysis incorporating all the variables in this study, thus multiple analyses were completed separately. The purpose of analyses in study 1 primarily involved comparing target versus target replacement ratings, and as such, the main analysis focused on this. A secondary focus was determining whether order effects were present. Additional analyses were conducted to explore the effects of participant gender, lineup gender, participant ethnicity and lineup ethnicity. However, these variables did not form part of the main focus of this research, and are typically reported on only briefly.
Table 5.1  
Study 1 Experimental Design

<table>
<thead>
<tr>
<th>Condition</th>
<th>Lineup Composition 1</th>
<th>Lineup Composition 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Lineup 1</td>
<td>Lineup 2</td>
</tr>
<tr>
<td>Male Caucasian</td>
<td>Present</td>
<td>Present</td>
</tr>
<tr>
<td>Female Caucasian</td>
<td>Early</td>
<td>Early</td>
</tr>
<tr>
<td>Male Asian</td>
<td>Present</td>
<td>Present</td>
</tr>
<tr>
<td>Female Asian</td>
<td>Early</td>
<td>Early</td>
</tr>
<tr>
<td>Male Asian</td>
<td>Present</td>
<td>Present</td>
</tr>
<tr>
<td>Female Asian</td>
<td>Late</td>
<td>Late</td>
</tr>
<tr>
<td>Male Asian</td>
<td>Absent</td>
<td>Absent</td>
</tr>
<tr>
<td>Female Asian</td>
<td>Late</td>
<td>Late</td>
</tr>
</tbody>
</table>

**Materials**

Photographs of lineup members were selected from the Multi-Pie Faces Database, which contains passport style photographs of Caucasian male, Caucasian female, Asian male and Asian female faces. Photos were 100x149 pixels. A target was selected for each lineup and then eight other photographs were selected on the basis of being visually similar to the target. For instance, if the target was a blond Caucasian male with short hair, the foil photographs were selected to match these features. Photographs containing distinctive traits such as glasses or facial hair were excluded from the lineups. Two different photographs of the target were used that were taken at different time points. The target photograph initially presented to participants differed in terms of background and clothing to that presented in the lineup. A total of 10 lineups were constructed which included two practice lineups. Photographs were selected for a distractor task on the basis of being highly dissimilar to the target. For instance, if the target was a young Asian female, the distractor photograph was of an older Caucasian male. See Appendix E for an example of a SIML similarity array employed in this study.

**Procedure**

Participants were directed to an online website where they were presented with a brief description of the study and the informed consent page (see Appendix A). Participants were then advised that they would view 10 lineups, of which the first
two were practice lineups. They were informed that the target may or may not be included in the photographs shown, and that the target’s appearance and/or clothing may have changed from the original photograph. Instructions were as follows:

*For each photographic lineup, you will be shown a photograph of a target person for five seconds.*

After the participant clicked ‘Next’ the target photograph was presented for five seconds. Participants were then advised:

*You will now be shown a photograph of a different person for two seconds. You will be asked to estimate this person’s age and to indicate how confident you are that your estimation is accurate.*

After the participant clicked ‘Next’ the distractor photograph was presented for two seconds. Participants estimated the age of the person and rated how confident they were in their estimation on five-point scale from ‘Not at all confident’ to ‘Extremely confident’. The distractor task, which entailed the photo presented for two seconds, age estimation and confidence rating, was included for two reasons. The first was to reduce the likelihood of ceiling effects, e.g. to reduce the likelihood of all participants correctly identifying the target. The second was to mimic in part the passage of time and memory interference that occurs in the real world where other faces are observed between an initial observation and subsequent identification.

Participants were then advised that they would view photographs of eight persons shown together, and be asked to look at each photograph and in their own time indicate how similar each person was to their memory of the target person. The eight photographs were then presented simultaneously with the instruction:

*Moving from left to right, please rate how similar each person is to your memory of the target person by entering a number from 0 to 100 in the corresponding box, where 0 represents ‘not at all similar’ and 100 represents ‘extremely similar’.*

This process was replicated for all the lineups. After completing both practice lineups participants were informed they would now commence the main study. The screenshots of this procedure can be viewing in Appendix F.

**Results**

Two measures of similarity are reported. The first and most important was the mean proportion in which the target or target replacement was rated most
similar to a participant’s memory of the target (e.g. received the highest rating in the lineup). This involved transforming ratings data into binary outcomes regarding whether the target was rated most similar (yes = 1) or not (no = 0). Whether the target was rated most similar or not was summed across the four target present lineups, providing a score between 0-4 for each participant. This process was also followed with regards to whether the target replacement was rated most similar across the four target absent lineups, providing a similar target replacement score between 0-4 for each participant. Thus the main analysis was completed on mean proportion scores of between 0-4. Secondary analyses were completed relating to specific manipulations of interest, and these often included a dependent variable score of between 0-2. For instance, if lineup gender was included in the analysis the dependent variable was between 0-2, as each participant only completed two male target present and two female target present lineups. The second measure, which is reported only infrequently, constituted the average ratings of targets, target replacements and foils. The proportion in which the target/target replacement was rated most similar was the preferred measure for three reasons: it was easily transformed into a binary outcome (e.g. target rated most similar Yes/No) making it more comparable to identification data; average ratings were limited by large standard deviations resulting from the 0-100% scale; and unequal group sizes meant that proportions were easier to interpret compared to frequencies.

Analyses entailed the comparison of mean proportions with repeated measures ANOVAs and t-tests using a probability level of $p < .05$. The assumption of homogeneity of variance was not violated unless specified. Non-parametric equivalents were also used to confirm significant results given that variables regularly failed the assumption of normality and were positively skewed. Four general analyses were conducted on similarity ratings exploring the impact of: (1) the presence/absence of the target, (2) the order of presentation, (3) participant and lineup gender, and (4) participant and lineup ethnicity.

(1) The impact of the presence/absence of the target

To explore the effect of the presence/absence of the target upon participant similarity ratings, variables were collapsed to allow for a comparison between the proportions of targets versus non-targets who received the highest rating in a lineup. The within participant independent variable was the presence/absence of the target. Between participant independent variables included participant gender and participant ethnicity. The two dependent variables constituted the proportion in which the target or target replacement received the highest rating (between 0-4).
Thus the analysis was 2 (target presence/absence) x 2 (participant gender) x 2 (participant ethnicity). The gender and ethnicity of the lineup were excluded as their inclusion reduced some cell sizes to less than 10 observations per cell. Furthermore, they were examined using separate analyses below.

Targets received the highest similarity rating in the lineup significantly more often ($M = 1.764$, $SE = .199$) than target replacements ($M = .959$, $SE = .156$): $F(1) = 7.826$, $p = .007$, 95% CI for the difference [.229 – 1.379]. Cohen’s effect size value ($d = .45$) suggested a small to moderate effect. This significant result was confirmed by a non-parametric Wilcoxon Signed-ranks Test [$W = 403.000$, $z = -3.829$, $p < .001$, with the effect size value ($d = .96$) indicating a large effect. None of the interactions were significant. Table 7.1 on page 59 shows the breakdown of similarity ratings for each of the eight lineups, including average ratings for targets and target replacements, as well as the proportion (in percentage terms) in which each received the highest rating, equal highest rating, and was out-rated by another lineup member.

(2) The impact of order of presentation

Targets/target replacements presented in positions two and three of a lineup constituted early presentation, and targets/target replacements presented in positions six and seven constituted late presentation. Participants either completed four early target present lineups, or four late target present lineups, but not both early and late target present lineups. This meant that order of presentation was a between participant variable. To explore the effect of order of presentation upon participant similarity ratings data were split and two separate independent samples t-tests were performed: the proportion of cases in which a lineup member received the highest similarity rating for (a) early versus late target ratings, and (b) early versus late target replacement ratings. The main effect for order was non-significant for both target present and target absent lineups.

(a) In the target present condition, while targets presented early ($M = 1.512$, $SD = 1.207$) tended to receive the highest rating less often than targets presented late ($M = 1.711$, $SD = 1.228$), this difference was not significant, $t(77) = -.724$, $p = .471$.

(b) In the target absent condition, similar looking target replacements presented early also tended to receive the highest rating slightly less often ($M = .711$, $SD = .835$) than those presented late ($M = .902$, $SD = 1.114$). Again, this difference was not significant, $t(77) = .861$, $p = .392$. 
(3) The impact of gender

To explore the impact of gender, participant’s ratings were summed to produce one male target rating, one female target rating, one male target replacement rating, and one female target replacement rating per participant. The dependent variable constituted the proportion of cases in which the target/target replacement was rated most similar. This was a score between 0-2 as the four target present lineups completed by each participant entailed two male lineups and two female lineups. The analysis was run on both target present and absent ratings separately, meaning that two repeated measures ANOVAs were conducted. For both analyses lineup gender was the within participant variable, and participant gender (Male N = 20, Female N = 59) was the between participant variable: 2 (lineup gender) x 2 (participant gender). Interpretations of findings based on gender were made with caution as there were fewer male compared to female participants. Even with participant ethnicity removed as a within participant variable, some male cell sizes were reduced to nine observations per cell. No significant results involving participant gender, lineup gender, or the interaction between the two were observed.

Gender of Participant

Male participants tended to rate targets most similar (M = .825, SE = .137) slightly more often than female participants (M = .763, SE = .080), but the difference did not approach significance: F(1) = .155, p = .695. Male participants tended to rate target replacements most similar (M = .500, SE = .103) more often than female participants (M = .339, SE = .060), and whilst more pronounced the difference failed to reach significance, F(1) = 1.838, p = .179.

Table 5.2 shows the mean ratings for targets, target replacements, target present foils and target absent foils according to participant gender. Males tended to provide higher average ratings of targets, target replacements and foils, compared to female participants. However, repeated measures ANOVAs with average ratings of foils as the dependent variable indicated that none of these differences approached significance.
Table 5.2
Overall mean ratings (0-100) according to participant gender

<table>
<thead>
<tr>
<th>Participant Gender</th>
<th>Mean Rating [Std. Dev.]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Target (TP)</td>
</tr>
<tr>
<td>Male (N=20)</td>
<td>46.2 [25.0]</td>
</tr>
</tbody>
</table>

Gender of lineup.

Male targets tended to be rated most similar ($M = .832$, $SE = .100$) more often than female targets ($M = .756$, $SE = .093$), but the difference did not approach significance: $F(1) = .479$, $p = .491$. In the target absent condition, female target replacements tended to be rated most similar ($M = .504$, $SE = .080$) more often than male target replacements ($M = .335$, $SE = .068$), and whilst the difference was more pronounced, it did not reach the level of significance: $F(1) = 3.587$, $p = .062$.

Interaction between gender of participant and gender of lineup.

There was no significant interaction between the gender of participant and gender of lineup with regards to ratings of targets [$F(1) = 1.307$, $p = .257$] or target replacements [$F(1) = .594$, $p = .443$].

(4) The impact of ethnicity

To explore the impact of ethnicity, participant’s ratings were collapsed to produce one Caucasian target rating, one Asian target rating, one Caucasian target replacement rating, and one Asian target replacement rating per participant. The dependent variable was the proportion of cases in which the target/target replacement was rated most similar (between 0-2). The analysis was run on both target present and absent ratings separately, meaning that two repeated measures ANOVAs were conducted. For both analyses lineup ethnicity was a within participant variable, and participant ethnicity (Caucasian N = 43, Asian N = 22) was a between participant variable: 2 (lineup ethnicity) x 2 (participant ethnicity). Thirteen participants who identified as ‘Other’ race were excluded from this analysis due to their small sample size and to provide a clearer picture of the interaction between Caucasian and Asian participants and lineups. With the target replacement analysis, one variable violated the assumption of homogeneity of variance. No significant
results were observed for participant ethnicity, lineup ethnicity, or the interaction between them.

**Participant Ethnicity.**

Asian participants tended to rate targets most similar \( (M = .886, \ SE = .132) \) more often than Caucasian participants \( (M = .756, \ SE = .094) \), but the difference did not approach significance: \( F(1) = .647, \ p = .424 \). Caucasian participants tended to rate target replacements most similar \( (M = .442, \ SE = .074) \) more often than Asian participants \( (M = .409, \ SE = .104) \), but again the difference did not approach significance: \( F(1) = .066, \ p = .798 \).

**Lineup Ethnicity.**

Asian targets tended to be rated most similar slightly more often \( (M = .873, \ SE = .101) \) compared to Caucasian targets \( (M = .769, \ SE = .104) \), however, this difference did not approach significance: \( F(1) = .676, \ p = .414 \). There was also a tendency for Asian target replacements to be rated most similar more often \( (M = .516, \ SE = .090) \) compared to Caucasian target replacements \( (M = .335, \ SE = .078) \), with the trend being more pronounced, but still failing to reach the level of significance: \( F(1) = 2.689, \ p = .106 \).

**Interaction between participant and lineup ethnicity.**

There was no significant relationship between the ethnicity of the participant and ethnicity of the lineup, \( F(1) = .215, \ p = .645 \). Both Caucasian and Asian participants tended to rate Asian targets most similar more often, which may reflect the impact of more distinct Asian targets overriding any tendency towards own-race bias.

### CHAPTER 6.

**Study 2**

**Method**

The main purpose of study 2 was to provide a basis for comparison with study 1. The aim was to determine whether the person rated most similar in study 1 was identified most often in study 2. Also of interest was determining whether more targets were correctly identified than rated most similar (comparative rates of accuracy with identifications and ratings). However, the comparison between rating and identification data itself occurs in Chapter 7 below. As such, this chapter describes the study 2 method and reports exclusively identification data. A
secondary aim of study 2 was to explore the impact of the independent variables on identification data (hits, misses, and false positives).

Participants

One-hundred-and-twenty-eight persons accessed the survey page, of which 52% completed the study. The vast majority (79%) dropped out in the practice lineup stage. Non-completers were excluded from analyses. There were likely two reasons for the lower dropout rate in study 2 compared to study 1. The task was easier, requiring one response per lineup (rather than eight in study 1), potentially leading to fewer persons dropping out prior to completion. Second, the single data point per lineup (compared to eight ratings data points) provided less scope for data to be missing.

The final sample therefore comprised a total of 67 participants who completed study 2. They were aged 16-70 years ($M = 28.8$, $SD = 11.7$), and were 64.2% female. Caucasian participants constituted 65.7% of the sample, with 19.4% being Asian and 14.9% of another race. Approximately 22% were born in Australia, with the remainder born overseas. Recruitment methods mirrored those of study 1.

Design

The study 2 design replicated that of study 1 (shown in Table 5.1) with one difference. Rather than completing similarity ratings for every lineup member, participants either selected one lineup member (e.g. the traditional identification process) or indicated that the target was not present. Similar to study 1 it was not possible to perform one analysis incorporating all the variables in this study, thus multiple analyses were completed separately.

Materials

Study 2 materials replicated those used in study 1.

Procedure

The procedure for study 2 replicated that in study 1 with the following exception. Study 1 participants were advised that they would view photographs of eight persons shown together (SIML) and complete ratings for the similarity between each lineup member and the participant’s memory of the target. Study 2 participants were advised that they would view photographs of eight persons shown together (SIML), and be asked to look at each and in their own time indicate whether they were able to identify the target person. The photographs were then presented
simultaneously and participants given the option of ticking a box underneath one photograph, or alternatively, ticking ‘Not present’. After completing both practice lineups participants were informed they would now commence the main study. See Appendix E for an example of a SIML identification array employed in this study.

Results

Dependent variables constituted hits, misses, and false positives (correct rejections were inferred from false positives). Any identification in a target absent lineup was treated as a false positive (e.g. no distinction was drawn between identifications of target replacements or foils). This was for several reasons. First, for the sake of simplicity: the focus of this research was not on false positives, and making this distinction would have added further complexity, which was considered unnecessary. Second, the major focus was on comparing identifications with ratings, and given that there is no false positive measure with regards to ratings, a comparison was not possible. This meant that only a cursory reporting of false positive data was necessary. Similarly, this research focused more on the target present condition, meaning that distinguishing between the two types of target absent false positives was not a priority.

Analyses entailed the comparison of means with repeated measures ANOVAs and t-tests using a probability level of $p < .05$. The assumption of homogeneity of variance was not violated unless specified. Non-parametric equivalents were also used to confirm significant results. ANOVAs were conducted on hits, false positives, and misses in the target present condition, as well as false positives (and by inference correct rejections) in the target absent condition. These entailed two observations per participant, giving a score between 0-2 per participant for each measure. Cell sizes were $\geq 10$. Analyses on gender and ethnicity were run separately, as running them together resulted in cell sizes of <10. Initial analyses indicated no significant interactions between participant gender and ethnicity. Twelve participants who identified as ‘Other’ race were excluded from analyses of ethnicity due to their small sample size and to provide a clearer picture of the interaction between Caucasian and Asian participants and lineups. Three general analyses were conducted on identification data exploring the impact of: (1) order of

---

4 There was no analysis corresponding with that in study 1 that involved scores between 0-4, as the target present/absent manipulation completed with regards to ratings was not meaningful with regards to identifications: e.g. comparing hits in the target present and target absent conditions was not possible.
presentation, (2) participant and lineup gender, and (3) participant and lineup ethnicity.

(1) The impact of order of presentation

Targets/target replacements presented in positions two and three of a lineup constituted early presentation, and targets/target replacements presented in positions six and seven constituted late presentation. Order of presentation (early vs. late) and participant gender were between participant variables, and lineup gender was a within participant variable. Thus, this repeated measures ANOVA was 2 (early vs. late presentation) x 2 (participant gender) x 2 (lineup gender). Separate ANOVAs were completed for the different dependent variables of hits, target present false positives, misses, and target absent false positives.

Target Present Hits.

The main effect of order was not significant, with hits being roughly equivalent between early ($M = .900, SE = .102$) and late ($M = .961, SE = .103$) presentation of the target: $F(1) = .181, p = .672$. No interactions involving order were significant. When this analysis was repeated with participant ethnicity replacing gender as the between participant variable, no interactions were significant.

Target Present False Positives.

The above analysis was repeated with target present false positives as the dependent variable. Levene’s test of equality of error variances was violated for the female lineup dependent variable. There was a tendency for more false positives to occur when the target was presented late ($M = .381, SE = .083$) compared to early ($M = .197, SE = .083$), though this difference was not significant: $F(1) = 2.462, p = .122$. No interactions involving order were significant, and this remained the case when participant ethnicity replaced gender as a between participant variable.

Target Present Misses.

The above analysis was repeated with target present misses as the dependent variable. Early presentation resulted in more misses ($M = .903, SE = .110$) compared to late presentation of the target ($M = .658, SE = .111$), though this was not significant: $F(1) = 2.491, p = .120$. No interactions involving order were significant, and this remained the case when participant ethnicity replaced gender as a between participant variable.
**Target Absent False Positives (and Correct Rejections).**

The above analysis was repeated with regards to target absent false positives. When the target is absent a false positive can refer to either the target replacement or a foil being selected. However, in this analysis any selection in target absent lineups was treated as a false positive. Slightly more false positives occurred when the target replacement appeared late ($M = .798, SE = .104$) compared to early ($M = .720, SE = .103$), though the difference did not approach significance: $F(1) = .283, p = .597$. No interactions involving order were significant, and this remained the case when participant ethnicity replaced gender as a between participant variable. Target absent correct rejections are not reported as they inversely mirror target absent false positives. For instance, if the mean false positive rate was .900 (out of 0-2), then logically, the mean correct rejection rate was 1.100 for the same condition. Thus, the results for target absent correct rejections are not repeated.

(2) **The impact of gender**

The 2 (early vs. late presentation) x 2 (participant gender) x 2 (lineup gender) repeated measures ANOVAs completed on the four dependent variable measures that are outlined above also provided the results reported here.

**Target Present Hits.**

There was a significant main effect for participant gender, with males correctly identifying the target ($M = 1.083, SE = .116$) more often compared to females ($M = .779, SE = .087$): $F(1) = 4.437, p = .039, d = .37$. This was confirmed by a non-parametric independent samples Mann-Whitney U test: $U = 351.500, z = -2.228, p = .026, r = .27$. There was no main effect for lineup gender [$F(1) = .316, p = .576$], with male ($M = .958, SE = .092$) and female ($M = .903, SE = .082$) targets being identified at roughly equivalent rates. There was no significant interaction between lineup gender and participant gender: $F(1) = .092, p = .762$.

**Target Present False Positives.**

There were no significant main effects or interactions involving participant and lineup gender for target present false positives. Female participants tended to make more false positives ($M = .328, SE = .070$) compared to males ($M = .250, SE = .094$), but the difference did not approach significance: $F(1) = .442, p = .509$. More false positives were made of female lineup members ($M = .333, SE = .065$) than male lineup members ($M = .245, SE = .068$), but the difference was not significant:
F(1) = 1.974, p = .165. The interaction between participant and lineup gender did not approach significance: F(1) = .006, p = .938.

**Target Present Misses.**

There were no significant main effects or interactions involving participant and lineup gender for target present misses. Male participants tended to make fewer misses (M = .667, SE = .125) compared to females (M = .894, SE = .093), but this difference was not significant: F(1) = 2.142, p = .148. Male and female targets were missed at roughly equivalent rates (male target M = .798, SE = .097; female target M = .764, SE = .091): F(1) = .106, p = .745.

**Target Absent False Positives (and Correct Rejections).**

No main effects or interactions involving participant or lineup gender were significant for target absent false positives. There was a non-significant tendency for male participants to make fewer false positives (M = .667, SE = .117) compared to females (M = .851, SE = .088): F(1) = 1.576, p = .214. Whilst non-significant, more false positives were made in relation to female lineups (M = .829, SE = .101) than male lineups (M = .688, SE = .088): F(1) = 1.399, p = .241.

**(3) The impact of ethnicity**

This analysis replicated that above, but with lineup ethnicity replacing lineup gender as a within variable, and participant ethnicity replacing participant gender as a between variable. Thus, the repeated measures ANOVA was 2 (early vs. late presentation) x 2 (participant ethnicity) x 2 (lineup ethnicity). The impact of ethnicity was interpreted with caution given the imbalance between Caucasian (N = 42) and Asian (N = 13) participants in the sample.

**Target Present Hits.**

There was a significant main effect for lineup ethnicity, with Asian targets being identified more often (M = 1.160, SE = .128) than Caucasian targets (M = .523, SE = .120): F(1) = 16.390, p = <.001, d = .71. This was confirmed by a non-parametric Wilcoxon Signed-ranks Test: z = -3.154, p = .002, r = .39. The main effect of participant ethnicity was not significant, though Caucasian participants tended to correctly identify the target more often (M = .927, SE = .092) than Asian participants (M = .756, SE = .169): F(1) = .787, p = .379. The interaction between lineup and participant ethnicity approached significance: F(1) = 3.650, p = .062. Caucasian participants identified Caucasian targets (M = .759, SE = .115) less often than Asian targets (M = .1.095, SE = .122). Asian participants also identified
Caucasian targets ($M = .288$, $SE = .212$) less often than Asian targets ($M = 1.225$, $SE = .225$), though the difference was more pronounced.

**Target Present False Positives.**

Levene’s test of equality of error variances was violated for the Asian lineup false positive dependent variable. Box’s Test of equality of covariance matrices was violated by a small margin ($p = .045$). No main effects or interactions involving ethnicity were significant for target present false positives. There was little difference between false positive rates for Caucasian ($M = .317$, $SE = .074$) and Asian ($M = .306$, $SE = .137$) participants: $F(1) = .005$, $p = .945$. There was a greater rate of false positives for Caucasian ($M = .407$, $SE = .101$) compared to Asian ($M = .216$, $SE = .086$) lineups, and this difference approached the level of significance: $F(1) = 3.414$, $p = .070$. Regarding the interaction between participant and lineup ethnicity, Caucasian participant false positive rates were equivalent across lineup ethnicity (Caucasian lineups $M = .327$, $SE = .096$; Asian lineups $M = .307$, $SE = .082$), whereas Asian participant false positive rates were greater for Caucasian than Asian lineups (Caucasian lineups $M = .488$, $SE = .177$; Asian ($M = .125$, $SE = .151$). However, this trend was not significant: $F(1) = 2.723$, $p = .105$.

**Target Present Misses.**

The main effect of lineup ethnicity was significant, with Caucasian targets being missed more often ($M = 1.069$, $SE = .125$) compared to Asian targets ($M = .624$, $SE = .122$): $F(1) = 8.972$, $p = .004$, $d = .52$. This was confirmed by a non-parametric related samples Wilcoxon Signed Ranks Test: $W = 117.000$, $z = -2.924$, $p = .003$, $r = .36$. The main effect of participant ethnicity was not significant, though Asian participants missed at a higher rate ($M = .937$, $SE = .174$) than Caucasian participants ($M = .756$, $SE = .094$): $F(1) = .848$, $p = .361$. There was no significant interaction between participant and lineup ethnicity: $F(1) = .759$, $p = .388$.

**Target Absent False Positives (and Correct Rejections).**

Levene’s test of equality of error variances was violated for the Asian lineup false positive dependent variable. The main effect for lineup ethnicity was significant. More false positives were made in Caucasian lineups ($M = .997$, $SE = .119$) than Asian lineups ($M = .611$, $SE = .114$): $F(1) = 7.251$, $p = .010$, $d = .47$. A non-parametric related samples Wilcoxon Signed Ranks Test failed to confirm this result as significant: $W = 158.500$, $Z = -1.877$, $p = .061$, $r = .25$. The main effect of participant ethnicity was not significant, with false positives occurring equivalently across Caucasian ($M = .795$, $SE = .087$) and Asian ($M = .813$, $SE = .161$).
participants: $F(1) = .009, p = .926$. The interaction between participant and lineup ethnicity was significant: $F(1) = 4.778, p = .033, d = .38$. This result was driven by Asian participants making more false positives in Caucasian ($M = 1.162, SE = .209$) compared to Asian ($M = .463, SE = .200$) lineups. Caucasian participants made false positives more evenly across both lineup ethnicities (Caucasian lineup $M = .832, SE = .113$; Asian lineup $M = .759, SE = .108$). When only Asian participants were included in the analysis, a non-parametric related samples Wilcoxon Signed Ranks Test found that the difference between target absent false positives across lineup ethnicity was significant: $W = .000, z = -2.640, p = .008, r = .73$. However, there were only 13 participants in this analysis.

CHAPTER 7.

Comparing Study 1 & Study 2

Results (Studies 1 & 2)

The comparison of similarity ratings and identification data has not been performed previously, and normal statistical procedures were inappropriate for comparing the two types of data. As such, only a descriptive comparison of these results is provided. Table 7.1 shows the proportions for similarity ratings and identifications for lineups 1-8. Overall the target was rated most similar 40% and identified 44% of the time. While this descriptive analyses compares highest rating data with identification data, it should be noted that a target being rated most similar did not necessarily equate with a hit. This is because it could not be determined that the highest rating would have resulted in an identification, as the option of rejecting the lineup was not provided. Similarly, a foil receiving the highest rating did not necessarily equate with a false positive. The focus of comparisons was between ratings and hits (e.g. rather than misses or false positives). This is because ratings better represent the upper threshold of similarity, and not the lower threshold of dissimilarity involved in rejecting a lineup. Thus, the relationship between ratings and identifications was expected to be strongest in relation to hits. Further, it was not possible to examine the relationship between correct rejections and ratings, as there was no equivalent to correct rejections in the ratings condition.

The most basic question in assessing the relationship between similarity ratings and identifications included asking ‘was the person rated most similar the most identified?’ The answer was that the target was rated most similar most often across all eight lineups, and was also identified most often. However, this was a
Table 7.1

Proportions of similarity ratings (study 1) and identifications (study 2) according to lineups 1-8 for target present and absent conditions.

<table>
<thead>
<tr>
<th></th>
<th>L1</th>
<th>L2</th>
<th>L3</th>
<th>L4</th>
<th>L5</th>
<th>L6</th>
<th>L7</th>
<th>L8</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Target Present</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Similarity Ratings</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average Rating</td>
<td>36</td>
<td>27</td>
<td>34</td>
<td>39</td>
<td>54</td>
<td>40</td>
<td>47</td>
<td>57</td>
<td>42</td>
</tr>
<tr>
<td>Highest Rating %</td>
<td>42</td>
<td>18</td>
<td>34</td>
<td>34</td>
<td>49</td>
<td>39</td>
<td>51</td>
<td>51</td>
<td>40</td>
</tr>
<tr>
<td>Equal Highest Rating %</td>
<td>13</td>
<td>0</td>
<td>5</td>
<td>18</td>
<td>12</td>
<td>20</td>
<td>7</td>
<td>22</td>
<td>12</td>
</tr>
<tr>
<td>Out-rated %</td>
<td>29</td>
<td>61</td>
<td>47</td>
<td>40</td>
<td>32</td>
<td>29</td>
<td>27</td>
<td>17</td>
<td>35</td>
</tr>
<tr>
<td>All Zero Responses %</td>
<td>16</td>
<td>21</td>
<td>13</td>
<td>8</td>
<td>7</td>
<td>12</td>
<td>15</td>
<td>10</td>
<td>13</td>
</tr>
<tr>
<td><strong>Identifications</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>% Targets identified</td>
<td>40</td>
<td>23</td>
<td>53</td>
<td>63</td>
<td>35</td>
<td>30</td>
<td>54</td>
<td>57</td>
<td>44</td>
</tr>
<tr>
<td>% False positives</td>
<td>0</td>
<td>20</td>
<td>17</td>
<td>10</td>
<td>22</td>
<td>27</td>
<td>11</td>
<td>11</td>
<td>15</td>
</tr>
<tr>
<td>% Misses</td>
<td>60</td>
<td>57</td>
<td>30</td>
<td>27</td>
<td>43</td>
<td>43</td>
<td>35</td>
<td>32</td>
<td>41</td>
</tr>
<tr>
<td><strong>Target Absent</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Similarity Ratings</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average Rating</td>
<td>24</td>
<td>30</td>
<td>31</td>
<td>37</td>
<td>23</td>
<td>32</td>
<td>22</td>
<td>23</td>
<td>28</td>
</tr>
<tr>
<td>Highest Rating %</td>
<td>10</td>
<td>14</td>
<td>20</td>
<td>29</td>
<td>18</td>
<td>34</td>
<td>11</td>
<td>26</td>
<td>20</td>
</tr>
<tr>
<td>Equal Highest Rating %</td>
<td>7</td>
<td>12</td>
<td>5</td>
<td>10</td>
<td>11</td>
<td>8</td>
<td>8</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>Out-rated %</td>
<td>59</td>
<td>59</td>
<td>59</td>
<td>51</td>
<td>58</td>
<td>42</td>
<td>58</td>
<td>55</td>
<td>55</td>
</tr>
<tr>
<td>All Zero Responses %</td>
<td>24</td>
<td>15</td>
<td>17</td>
<td>10</td>
<td>13</td>
<td>16</td>
<td>24</td>
<td>13</td>
<td>16</td>
</tr>
<tr>
<td><strong>Identifications</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>% Target-Replace identified</td>
<td>24</td>
<td>22</td>
<td>11</td>
<td>24</td>
<td>13</td>
<td>30</td>
<td>7</td>
<td>7</td>
<td>17</td>
</tr>
<tr>
<td>% False positives</td>
<td>24</td>
<td>22</td>
<td>30</td>
<td>11</td>
<td>20</td>
<td>23</td>
<td>10</td>
<td>33</td>
<td>22</td>
</tr>
<tr>
<td>% Misses</td>
<td>51</td>
<td>57</td>
<td>59</td>
<td>65</td>
<td>67</td>
<td>47</td>
<td>83</td>
<td>60</td>
<td>61</td>
</tr>
</tbody>
</table>

Note. This table illustrates the mean similarity ratings, the proportion of cases in which the target/target replacement received the highest similarity rating in the lineup, the equal highest rating, or was out-rated by another lineup member, and the proportion of cases where participants rated all members of a lineup as having zero similarity, for lineups 1-8. The identification data shows the proportion of targets, false positives (collapsed across both target replacements and foils), and misses.

Differentiation & Match: A method for exploring the relationship between similarity ratings and identifications

One way of conceptualising the relationship between similarity ratings and identifications included the use of ‘differentiation’ and ‘match’. Differentiation pertained to similarity ratings, and described whether one person in the lineup stood out. It was determined by the difference between the lineup member who was rated most similar most often and the lineup member who was rated most similar second
most often. Thus it was based on collective rather than individual data and did not involve average ratings. Low differentiation occurred when there was no clear person rated most similar more often than any other; thus the difference between the two persons rated most similar was minimal (see Table 7.2 example 2 for an illustration of Low-D – person 3 and 5 are rated most similar equally often). High differentiation occurred when one person was rated most similar clearly more often than any other lineup member, producing a larger difference between the two most highly rated lineup members (see Table 7.2 example 1 for an illustration of High-D – person 3 is rated most similar clearly most often).

Differentiation was useful as it provided an indication of the conditions in which a stronger relationship between similarity ratings and identifications would be expected. For instance, if no one lineup member was rated most similar most often (Low-D) then no one person would be expected to be identified most often, and the relationship between ratings and identifications would be diminished. Conversely, when one person was perceived as most similar more often (High-D), then it was expected that they would be more likely to be identified more often. This correspondence between ratings and identifications was termed ‘match’. Where differentiation described the relationship between the two highest similarity ratings, ‘match’ described whether the lineup member rated most similar was in fact identified most often in the corresponding condition. Thus, match described how well the two types of data agreed, and provided a useful measure of the strength of the relationship between similarity and identity. It was expected that if similarity ratings underpinned identifications, then a stronger degree of match should exist for lineups where one member was highly differentiated: a person that was High-D based on ratings should be identified most often (High-M). Naturally, if no lineup member stood out, then it would be more difficult to observe correspondence between the data, and the relationship between ratings and identifications would be obscured. While typically it was the target that was given the highest rating most often, using differentiation, it was irrelevant whether the person rated most similar was the target, target replacement or foil - of interest was comparing whether the same person was identified most often in the corresponding condition. Thus, three things were of interest: overall levels of differentiation, overall levels of match, and how often differentiation predicted match. The latter allowed determination of whether the relationship between ratings and identifications was stronger for High-D compared to Low-D lineups.
Table 7.2

Example of a lineup demonstrating High-D and High-M on the left and Low-D and Low-M on the right. The proportion in which a lineup member was rated most similar (HR%) and/or identified is plotted on the Y-axis, with the lineup persons (1-8, or 9 = 'not present') plotted on the X-axis.

Example 1: High-D & High-M
Example 2: Low-D & Low-M

For the purpose of this study, differentiation was operationalised as follows:

- Low differentiation (Low-D) = less than or equal to 10% difference between the first and second persons rated most similar; Moderate differentiation (Mod-D) = between 10.1 - 20% difference; and High differentiation (High-D) = when the difference between the two exceeded 20%. Using the first example in Table 7.2, person 3 was High-D as they were rated most similar 32% of the time, and the next person was rated most similar 10% of the time, with the difference exceeding 20%. The lineup would be Low-D if the next most highly rated person was rated most similar equal to or more than 22% of the time (32% minus 10% = 22%).

Level of match was

---

5 Differentiation can be defined in either relative or absolute terms. Relative differentiation involves taking a percentage of a percentage, e.g. if the person rated most highly most often was rated most highly 45% of the time, then for the lineup to be High-D the next person would need to be rated most highly less than 33.75% of the time (20% of 45% = 11.25, 45%
Identity Crisis in Identification Evidence

operationalised as follows. A high degree of match (High-M) occurred if the person rated most similar most often was the most commonly identified (see Table 7.2 example 1, where person 3 was rated most similar most often and also identified most often). Moderate match (Mod-M) occurred when the person rated overall most similar was the second most commonly identified, or if the person most commonly identified was rated most similar the second most often. Low match (Low-M) occurred when the person rated overall most similar in a lineup was not the first or second most commonly identified (see Table 7.2 example 2).

Given that every lineup included four conditions (target present early presentation, target present late, target absent early and target absent late), these conditions were explored separately to better examine differentiation and match. This allowed a comparison of the 32 lineup conditions presented in Table 5.1. As shown in Table 7.3, overall level of differentiation was high in 13/32 cases, moderate in 7/32, and low in 12/32. Overall level of match was high in 17/32 cases, moderate in 6/32 cases, and low in 9/32 cases. Level of differentiation accurately predicted the level of match in 22 cases (highlighted in Table 7.3). Of the 10 lineups where differentiation and match did not agree (not highlighted in Table 7.3), there was a discrepancy of one category (e.g. Low-D and Mod-M) in nine cases (9/32), and an extreme discrepancy (e.g. Low-D and High-M) in one case (1/32). See Table 7.4 for an example of this extreme discrepancy.

Table 7.3

<table>
<thead>
<tr>
<th>Level of Differentiation</th>
<th>Level of Match</th>
<th>Total</th>
<th>Predictability%</th>
</tr>
</thead>
<tbody>
<tr>
<td>High-D</td>
<td>13</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Mod-D</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Low-D</td>
<td>1</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>Total</td>
<td>17</td>
<td>6</td>
<td>9</td>
</tr>
</tbody>
</table>

minus 11.25% = 33.75). Absolute differentiation applies the same difference to all comparisons, e.g. with a 45% highest rating, for the lineup to be High-D the next highest rating would need to be less than 25% (45% minus 20% = 25%). The absolute definition was employed here as it was simpler and more easily understood. However, the relative definition better accounts for the subjective nature of ratings - one person’s baseline rating may be highly different to another’s.
Table 7.4
Example of extreme discrepancy (no relationship between similarity ratings and identifications) with High-D and Low-M. The person rated most similar (6) is not the most or second most identified.

Example 3: High-D & Low-M

Apart from overall matching, one important question was ‘of all the lineups that were High-D, how many predicted High-M?’ It was expected that similarity ratings would best predict identifications when a person was highly distinctive. This would appear to reference similarity rather than dissimilarity processes. The answer was that High-D predicted High-M 100% of the time (or in 13/13 cases). Most of these High-D lineups (92.3% or 12/13) pertained to targets, with one relating to a target replacement. This meant that when one person in the lineup was perceived to be notably more similar to the target than others, then this same person was identified more often in the corresponding condition, and provides evidence for a relationship between ratings and identifications. Mod-D predicted High-M in 3/7 cases, and Low-D predicted High-M in 1/7 cases, showing a clear relationship between increased similarity and increased likelihood of that person being identified. In comparison, Mod-D predicted Mod-M only 28.6% of the time (or 2/7 cases), and Low-D predicted Low-M 58.3% of the time (or in 7/12 cases). Thus, when differentiation decreased, the relationship between ratings and identifications also decreased.
When differentiation and match were examined according to the presence/absence of the target, it was discovered that differentiation predicted match in 12/16 cases when the target was present, all of which involved High-D predicting High-M. In comparison, when the target was absent differentiation predicted match in 10/16 cases, most of which involved Low-D predicting Low-M (7/10). This may indicate that in target present lineups the relationship between ratings and identifications was based on similarity processes, but in target absent lineups was based on dissimilarity processes.

That targets were more differentiated compared to target replacements or foils indicates that ratings were both capable of, and useful for, distinguishing between faces that had been presented earlier and those that had not. Furthermore, ratings were more diagnostic of identifications when differentiation increased. It is likely that even when a lineup member is differentiated from others, a certain threshold of similarity may be necessary for that person to be identified. It is important to acknowledge that a lineup member may be differentiated due to either being highly dissimilar to other lineup members (perceptual dissimilarity), or highly similar to a person’s memory of the target (ecphoric similarity). Similarly, a lineup member may have high ecphoric similarity, but be Low-D due to also being perceptually similar to the remaining lineup members. Increased ecphoric similarity would be more likely to result in an identification than perceptual dissimilarity. However, in the current study it was not possible to determine which process was relied upon for any given judgment.

**Graphic representation of the 8 lineups**

The 32 conditions were collapsed to produce 8 lineup graphs (see Appendix B), which demonstrate visually the relationship between the similarity ratings and identifications for the eight lineups. These graphs depict the relationship between all lineup members (including the target, target replacement and foils), which as evident in lineups 2 and 4, can demonstrate a noticeable level of symmetry between similarity ratings and identifications.

**Discussion (Studies 1 & 2)**

The general aim of this research was to explore the relationship between similarity ratings and identification data. This was considered the first step of determining whether ratings can provide an alternative to identifications as the method for obtaining and presenting eyewitness evidence. This study addressed the following four research questions: (1) Are similarity ratings related to identification
patterns? (2) Under what conditions are ratings most diagnostic of identifications? (3) Do similarity ratings provide a useful and viable alternative to identifications? (4) How do similarity ratings differ across the experimental conditions (target present/absent, early/late presentation, male/female lineups, and Caucasian/Asian lineups)?

(1) Are similarity ratings related to identification patterns?

Results provide preliminary evidence for a relationship between similarity ratings and identifications. The presence of a relationship was determined by asking the question ‘is the person rated most similar identified most often in the corresponding condition?’ When one lineup member was perceived to be on average notably more similar than others (e.g. highly differentiated), they were also identified more often than any other lineup member in 100% of cases. Furthermore, there was a high or moderate level of agreement (e.g. match) between the two types of data in 72% of cases. This meant that the person rated most similar was either the first or second most identified (or that the person most identified was rated first or second most similar).

Aside from determining whether a lineup member stood out in terms of ecphoric similarity (differentiation), and whether that same person was identified most often (match), another question was whether the level of differentiation predicted the level of match. This was a useful measure as it accounted for Low-D lineups where there was less expectation for match to be high; if no one person was rated clearly most similar, then it was considered less likely for them to be identified most often, thus match would be low in this context. The level of differentiation predicted match 69% of the time, providing support for the hypotheses that the level of agreement between ratings and identifications would be mediated by the amount of ecphoric similarity perceived.

(2) Under what conditions is the relationship strongest?

The relationship between similarity ratings and identifications appeared to become stronger under certain circumstances. Namely, similarity ratings were more diagnostic of identifications when differentiation increased. When a person was clearly rated most similar (highly differentiated) they were much more likely to be identified compared to a person that was rated second most similar or less (moderate or low differentiation). The relationship between similarity and identity may be driven by ecphoric similarity or perceptual similarity, or a combination of both. For instance, high differentiation may be a product of a lineup member
standing out from other lineup members (perceptual similarity), or of a lineup member approximating a witness’s memory of the target (ecphoric similarity). However, it is difficult to tease apart these two types of similarity in practice. It may be possible to focus exclusively on perceptual similarity by making the comparison between two present stimuli (e.g. present the target photo with the lineup). However, it is far more difficult (if not impossible) to eliminate the influence of perceptual factors and focus solely on ecphoric processes. This is because ecphoric judgments will always rely in part on perceptual processes. That said, it would be useful to be able to determine which process underpins a similarity judgment for two reasons; ecphoric processes will lead to greater accuracy (as they relate to the memory of the perpetrator rather than similarity between lineup members), and will also better correlate with identification decisions. This is because a lineup member may be highly differentiated based on perceptual processing (stand out), but not approach the threshold of ecphoric similarity necessary for an identification to occur.

This tendency for the relationship to be strongest when differentiation was high was evident in relation to target present and target absent conditions. Targets were rated as clearly most similar (high differentiation) 75% of the time (or in 12/16 lineups). In every one of these cases (100%) the same person was identified most often in the corresponding condition. In comparison, only 6% (or 1/16 lineups) of target absent lineups included a person who was clearly perceived as being most similar (high differentiation). That this target replacement was also identified most often provides strong evidence for a relationship between similarity and identity. While differentiation predicted match roughly equivalently across target present and absent lineups, target present lineups typically involved High-D predicting High-M, whereas target absent lineups involved Low-D predicting Low-M. This may indicate that target present lineups referenced similarity processes, whereas target absent lineups referenced more dissimilarity processes. That only one target absent lineup was High-D may also indicate that ecphoric processes underpinned similarity ratings.

Ratings were highly diagnostic of accuracy when differentiation was high, with 92.3% or 12/13 High-D lineups pertaining to targets (e.g. one High-D lineup involved the identification of a target replacement). However, how to utilise this presents a conundrum, as it is not possible in practice to determine whether a lineup includes a suspect or not. Baseline ratings may prove useful here. For instance, a person may be High-D but have a low rating overall in the context of a witness’s rating pattern. To understand this would involve creating a rating profile for a
Identity Crisis in Identification Evidence

witness based on them completing a number of standardised practice lineups. However, in the current study, the target replacement received a high average rating overall, suggesting that they were perceived to be highly similar to participant’s memory of the target. Thus, average ratings will not remedy this problem when ecphoric similarity is high.

Reasons for a disconnect between similarity and identity.

Whilst a relationship between similarity and identity was observed, it was not perfect. There are several reasons to expect some level of disconnect between similarity and identity. First, as indicated, the relationship will be reduced when ecphoric similarity is low. It will also be reduced when perceptual similarity is low; whilst identifications are more likely to be based on ecphoric judgments, they will also be influenced by perceptual processes – a person may be identified due to standing out from other lineup members (as well as approximating the target to some degree). Thus, if little similarity is perceived, then it is inevitable that it will be less likely to predict identifications. Second, even when similarity is perceived, it will depend on whether it relies on ecphoric or perceptual processes, with the former expected to relate more strongly to identifications. Third, it is unknown whether a participant’s judgments were based upon local (feature-based) or global (holistic) processes. A person identified by a participant relying on a holistic approach may not have been rated as more similar by a participant relying upon a feature-based approach, or vice versa. Even when two persons rely on a feature-based judgment, the features they rely most heavily upon to gauge similarity may differ, leading to a discrepancy in perceptions of similarity (and ratings). For instance, one person may focus on the similarity of the hair colour and style, whilst another may focus on the shape of the nose and mouth. Fourth, a disjuncture between similarity and identity exists in that a person may look visually similar to a target without being the same person. Thus an innocent person may be perceived to be highly similar, and consequently incorrectly identified. This was evident in the current study by one target replacement being rated as highly similar to participants’ memory of the target. This person was also identified most often, providing strong evidence for a relationship between similarity and identity. However, it also highlights the potential disconnect between ratings/identifications and actual identity, and therefore their capacity for inaccuracy.

These factors initially present as flaws in the similarity ratings process, in that ratings do not necessarily seamlessly mirror identifications. However, these factors are equally problematic with regards to identifications; e.g. a similar looking
foil may be identified. If anything, these limitations serve to highlight a substantial problem with the identification process; namely, it is not clear upon what basis identifications are made. In this sense identifications are comparable to an iceberg, of which the greater part lies beneath the surface, unobservable and unknown. Similarity ratings provide a means for a more informed mapping of the process.

(3) Do similarity ratings provide a useful and viable alternative to identification?

Did ratings distinguish between targets and foils, and targets and target replacements?

Results showed a significant difference between similarity ratings of targets and target replacements. Overall, targets were rated most similar at roughly twice the rate of target replacements. This suggests that ratings were useful for indexing participant perceptions of similarity and accurately discriminated between photos of targets and target replacements. Average ratings of targets were also typically two or three times greater than average ratings of foils, again suggesting that ratings were useful for discriminating between targets and foils.

The fact that the target was out-rated by another lineup member 35% of the time may partly reflect the difficulty of the task, which involved viewing a different target photo initially and in the lineup (with a different background). It is not possible to determine with the current design whether these ratings might reflect false positives in the identification condition, as the lineup may have been rejected in the later situation.

Were ratings diagnostic of accuracy?

Overall, slightly more targets were identified (44%) than rated most similar (40%). This finding is in contrast to Brewer et al. (2012) who found that making time-pressured confidence judgments regarding whether a face had been seen before was more accurate than identification decisions. However, a number of differences exist between the two studies; the current study utilised the simultaneous lineup, was not time-pressured, and employed similarity ratings. That identifications were slightly more diagnostic of accuracy might appear to be an argument for retaining the identification procedure; that there is something inherent to identifications that makes them more accurate. For instance, it might be that identifications in this study relied more upon global processes involving participants quickly zoning in on the lineup member that most represented their memory of the target. In contrast, the
ratings condition may have forced a more conscious and laborious process relying on feature-based processing.

With identifications there was no way of determining whether a selection was made with certainty or not – e.g. it could not be determined whether decisions were made with the response criterion operating around the threshold for making a decision or not. While some participants made an identification with the response criterion operating around this threshold (which may be a consequence of forcing a binary decision upon what actually occurs on a continuum), in the ratings condition these threshold decisions were better captured by the ‘equally highly rated’ measure. Using ratings, 12% of targets were rated equally highly as another lineup member. Correspondingly, one or more lineup members were rated equally as high as the target replacement in 8.2% of cases. This does not include situations where all members were rated at 0% similarity, which would inflate the previous figure.

Another measure captured by ratings was ‘zero responding.’ This referred to situations where a participant rated every lineup member within a target present lineup as 0% similar to their memory of the target. This occurred in 13% of cases. Given that no participant provided zero responses for every lineup, this may index an absence of perceived similarity rather than absence of effort. As such, this figure potentially reflects instances where participants had difficulty making eyewitness judgments. This might be due to difficulties with memory, attention or perceptual problems. Whatever the reason, the use of ratings allows this event to be detected. Using a traditional identification process this event would not be observed. Zero responding increased slightly in target absent lineups to 16.5% of cases.

**Advantages of using ratings**

Identifications do not allow for much interpretation of a witness’s decision. A simple binary outcome is presented to the jury who are unable to probe it. An exception to this is the United States where a witness’s confidence is presented to the jury as an indicator of accuracy. However, this is problematic as the link between confidence and accuracy is tenuous at best (Leippe, 1980). In comparison, ratings allow for at least three questions to determine the accuracy of a witness’s decision. These include (1) whether the suspect was rated most similar, (2) whether the suspect was provided with a high rating overall, and (3) how large was the difference between the highest and second highest rating (this third question is explored in study 3). As noted above, to interpret whether a witness’s highest rating is subjectively high or low (and therefore more indicative of greater accuracy) would
require obtaining further information on their subjective rating tendencies. Ratings also provided a more nuanced picture of eyewitness decisions including whether another person was rated equally highly.

Thus, results from this explorative analysis provide preliminary evidence that similarity ratings offer two benefits not provided by the traditional identification process. First, they provide a means for examining in more detail participant’s perceptions of the similarity between a previously seen target and members of a lineup. Thus, from a research perspective, similarity ratings allow for a more detailed analysis of eyewitness decisions. Second, similarity ratings allow for more information to be provided to jurors who can then engage with the evidence, and make a more informed decision regarding the evidence’s reliability.

(4) How do similarity ratings differ across the experimental conditions (target present/absent, early/late presentation, male/female lineups, Caucasian/Asian lineups)?

As noted, targets were rated most similar significantly more often than foils and target replacements. Average ratings of targets were also typically two or three times greater than average ratings of foils. There were no significant differences based on whether the target or target replacement was presented early or late, suggesting that the SIML procedure incorporating ratings was robust against order effects. The SIML was also robust against order effects in the identification condition. There were no significant differences in ratings based on the gender or ethnicity of participant, or gender or ethnicity of lineup. There was a non-significant tendency for male average ratings overall to be higher than those of females across ratings of targets, target replacements and foils. This suggests that the male participants in this study had a higher baseline for ratings compared to females. However, findings regarding gender and ethnicity were interpreted with caution given the small number of Asian and male participants.

Limitations and recommendations

Similarity ratings can be examined based on average ratings or whether the target or target replacement was rated most similar. One discovery made from this study was that there were three advantages to using the latter. First, the 100% scale employed here resulted in large standard deviations, which made the observation of significant results based on average ratings less likely. Second, such a large scale was unrealistically fine and may have created an illusion of precision; e.g. a person is unlikely to perceive a difference of 1%. Third, average ratings sometimes
obscured the relationship between similarity ratings and identifications, where the highest ratings measure appeared to bear it out. As such, future research on ratings should employ a smaller five, seven, nine, or 11-point Likert scale.

The current study’s experimenter selected targets and target replacements on the basis of their being visually similar. However, this process could be better objectified and operationalised by randomly selecting targets, and then having a pool of persons rate the similarity of a group of photos to the target. This would produce the target replacement (1st most highly rated), and other foils (the 2nd – 8th most highly rated). If multiple persons were rated as equally similar to the target, then they could be removed from the lineup, allowing for better control of the similarity within the lineup. This was addressed in study 3.

This study focused on ecphoric similarity, though the influence of perceptual factors could not be excluded. Future research should explore whether perceptual (visual) similarity is more or less strongly related to identification patterns than ecphoric similarity. This could involve presenting an additional target photo alongside the standard lineup to encourage perceptual rather than memory-based similarity judgments. For instance, one target photo would be presented above the lineup. Thus, using an eight person lineup, nine photos would be visible – the first target photo and the standard eight person lineup. Participants would then visually compare the target to each photo in the lineup, including comparing two photos of the target. Accordingly, visual similarity was included in study 3 to allow for a more nuanced examination of the relationship between similarity and identity. Whilst it is expected that ecphoric factors more closely underpin identifications, this remains unverified.

Perhaps the most significant debate in the eyewitness identification field at present is between the SIML and SEQL procedure, and determining which is more meritorious. This study included only the SIML procedure, thus a major focus in study 3 was to apply the similarity ratings process utilised here to the SEQL procedure. Whether a similar pattern of results would be obtained for both lineup procedures is of great interest. It is possible that the SEQL procedure would better encourage reliance on ecphoric processes by limiting the ability for inter-lineup (perceptual) comparisons. As noted, there are currently no published studies incorporating similarity ratings. Consequently, there was a need to complete further research on similarity ratings to include the above recommendations, as well as determine whether the findings reported here are replicable.
Summary of findings

Initial findings provide preliminary support for the hypothesis that similarity judgments are related to identification decisions, and this relationship was strongest when differentiation increased: When one lineup member was perceived to be on average notably more similar than others (e.g. highly differentiated), they were also identified more often than any other lineup member in 100% of cases. Given that similarity is an aspect of identity, rather than the reverse, it appears probable that similarity processes underpin identification processes rather than the other way around. If so, then the more nuanced and precise measure of similarity ratings should be employed in obtaining evidence rather than the current identification process. However, further research is required to determine this. Thus, differentiation proved to be a useful measure for gauging the relationship between similarity and identity. Targets were rated as clearly most similar (high differentiation) 75% of the time (or in 12/16 lineups). In every one of these cases (100%) the same person was identified most often in the corresponding condition. In comparison, only 6% (or 1/16 lineups) of target absent lineups included a person who was clearly perceived as being most similar (high differentiation).

Results reported here have demonstrated the utility of similarity ratings for obtaining and presenting more detailed evidence. Similarity ratings capture more information than identifications and provide an index of a witness’s certainty or uncertainty, appear to be accurate at a roughly equivalent rate to identifications, may relieve pressure on a witness during the initial identification stage, remove the illusion of certainty associated with identifications, provide more information to the jury, and return the decision making role to the jury. In future legal processes, a jury might be advised whether the suspect was rated most similar or not, and if they were, whether any other lineup members were rated at nearly equivalent levels. In cases where there are multiple persons rated highly, this is important information for the jury to consider. In cases where one person is clearly rated most similar, the (false) impression of certainty associated with identifications may be removed.

CHAPTER 8.

Study 3

The aim of study 3 was to continue to explore ratings in relation to additional conditions, and determine whether the results of study 1 would be replicated. The ratings in study 3 were compared to identifications in study 4. This allowed for
further exploration of the relationship between ratings and identifications, and whether ratings might present a viable alternative to identifications. Study 3 replicated study 1, but with a number of alterations. The following changes were made based on recommendations from studies 1 and 2: (1) the selection of targets was randomised and the process for the selection of lineup members operationalised, which is described in more detail below; (2) both simultaneous and sequential lineups were included to address the question of whether similarity ratings differ across the two lineup procedures; (3) whereas in study 1 ratings were based on ecphoric similarity, in study 3 a new ‘visual’ similarity task was added where ratings were made on the basis of perceptual similarity; (4) incorporated the term ‘resemblance’ rather than ‘similarity’ (this change was made due to the use of the term ‘resemblance’ in legislation – it was assumed that it was measuring the same concept of similarity explored in study 1); (5) employed a 7-point Likert-scale for resemblance ratings rather than the 0-100% scale; and (6) a simplified lineup design including only Caucasian male faces, where lineups 1, 2, 5 and 6 always included the target, and lineups 3, 4, 7 and 8 always included a target replacement, and where early presentation entailed position three, and late presentation position 6. The simplified design allowed for easier comparison between conditions in the analysis stage.

Method

Participants

Across the four similarity conditions 149 persons accessed the survey page, of which 58% completed the study. The majority dropped out before completing lineup 1. Overall completion rates were higher than study 1, which may reflect the task being easier due to the inclusion of exclusively male Caucasian lineups. Non-completers were excluded from analyses. The final sample therefore comprised a total of 90 participants who completed study 3 across the four similarity conditions. They were aged 15-66 years (\( M = 28.9, SD = 12.4 \)), and were 65.6% female. Caucasian participants constituted 63.3% of the sample. Recruitment processes mirrored study 1.

Design

Between participant variables included condition (memory/visual task), lineup procedure (SIML/SEQL), participant gender and participant ethnicity. Within participant variables included order of presentation (early/late) and target presence/absence. Thus, study 3 incorporated a 2 (condition) x 2 (lineup procedure)
Identity Crisis in Identification Evidence

x 2 (participant gender) x 2 (participant ethnicity) x 2 (order of presentation) x 2 (target presence/absence) mixed design. See Table 8.1 for a visual representation of the study design.

The newly introduced visual task entailed participants viewing a target face alongside the lineup. This meant that there were nine faces visible on the screen in the simultaneous visual condition, with the first target photo being presented above the remaining eight photos of the lineup. See Appendix E for an example of a SIML visual condition similarity rating array employed in this study. In the sequential condition, two photos were visible, with the target photo being presented above one other lineup photo. Thus participants saw two different photos of the target in target present lineups. The two target photos depicted the target in the same clothes, though were taken on different days and featured different backgrounds. Participants then rated the resemblance of each lineup member to the first target photo. Thus, it was assumed that these ratings were made on the basis of visual or perceptual similarity (how similar lineup members are with each other), rather than ecphoric similarity (how similar lineup members are with characteristics of the target in the witness’s memory). The visual condition was included as a way of examining in more detail the relationship between perceptual and ecphoric similarity and identifications. An important distinction regarding perceptual similarity is necessary here. In study 1 it exclusively referred to the similarity between lineup members and was an extraneous influence. In the study 3 visual condition it predominantly provided a measure of the visual similarity between the first target photo and remaining lineup members. Thus, in this latter context, it was a desired perceptual comparison that included the target, rather than referring merely to inter-lineup member comparisons as in study 1. Given that participants could compare each lineup photo to the target photo, it was unlikely that they had cause to revert to inter-lineup comparisons for visual cues. This type of perceptual similarity was expected to form a more basic building block of similarity upon which ecphoric judgments are founded. Thus it was predicted that perceptual similarity ratings in the visual condition would underpin ecphoric ratings and identifications in the memory-based condition.

The memory task replicated the standard condition employed in study 1, where the target photo was presented initially, and then similarity ratings were made on the basis of a participant’s memory of the target photo. As in study 1, the photo of the target presented to participants initially in the observation stage (or with the lineup in the visual condition) was a different photo of the target than the one...
presented in the lineup. This was the case across all conditions. These two target photos differed in terms of their background. They both depicted the person in the same clothing, from the same angle and in similar lighting conditions. The order in which participants completed lineups 1-8 was randomised. The order in which the faces appeared in the lineup was held constant. The latter was due to a limitation of the computer program employed, which prevented the order of photos being randomised in the SEQL.

Table 8.1
Study 3 Experimental Design

<table>
<thead>
<tr>
<th>Similarity Ratings</th>
<th>Lineup 1</th>
<th>Lineup 2</th>
<th>Lineup 3</th>
<th>Lineup 4</th>
<th>Lineup 5</th>
<th>Lineup 6</th>
<th>Lineup 7</th>
<th>Lineup 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIML Visual Task</td>
<td>Present / Early</td>
<td>Present / Late</td>
<td>Absent / Early</td>
<td>Absent / Late</td>
<td>Present / Early</td>
<td>Present / Late</td>
<td>Absent / Early</td>
<td>Absent / Late</td>
</tr>
<tr>
<td>SIML Memory Task</td>
<td>Present / Early</td>
<td>Present / Late</td>
<td>Absent / Early</td>
<td>Absent / Late</td>
<td>Present / Early</td>
<td>Present / Late</td>
<td>Absent / Early</td>
<td>Absent / Late</td>
</tr>
<tr>
<td>SEQL Visual Task</td>
<td>Present / Early</td>
<td>Present / Late</td>
<td>Absent / Early</td>
<td>Absent / Late</td>
<td>Present / Early</td>
<td>Present / Late</td>
<td>Absent / Early</td>
<td>Absent / Late</td>
</tr>
<tr>
<td>SEQL Memory Task</td>
<td>Present / Early</td>
<td>Present / Late</td>
<td>Absent / Early</td>
<td>Absent / Late</td>
<td>Present / Early</td>
<td>Present / Late</td>
<td>Absent / Early</td>
<td>Absent / Late</td>
</tr>
</tbody>
</table>

Note: All lineups exclusively included photos of Caucasian males. SIML indicates the simultaneous lineup, and SEQL indicates the sequential lineup procedure.

Materials

The lineup construction process differed substantially from studies 1 and 2, and involved a three-stage process. First, eight target photos were randomly selected from the same Multi-Pie Faces Database as study 1. Photos were of the same size and quality as study 1, and only male Caucasian faces were included. Unique faces, such as those including distinctive features including glasses or baldness were excluded from selection as targets. Second, for each target photo 30 Multi-Pie photos were randomly selected from the remaining photos. Six Caucasian members of a research team (four male) at Deakin University then rated these 30 photos in terms of their similarity to the target. Of the 30, the photo rated on average most similar to the target was selected as the target replacement for the four target absent lineups. Third, the next seven photos rated most similar were included as foils. Whilst unique faces (e.g. glasses/bald) were excluded from selection of targets they were not excluded from selection of foils. They were included on the basis that faces rated least similar to the target would be excluded from finalised lineups. A person with glasses may have been rated as more similar to the target than another without glasses. Thus, the inclusion of unique faces allowed the similarity process to be completely ratings driven. Any photo used in one finalised lineup was not used in
a subsequent lineup. This process created eight male Caucasian lineups with eight persons each. A ninth practice lineup was also included that was made up of female Caucasian faces.

**Procedure**

The memory task ratings procedure replicated that described in study 1, but with several changes. As noted, the term ‘similarity’ was replaced with ‘resemblance’, and participants were not instructed to “Move left to right” with regards to ratings. The instruction to “Move left to right” was removed as it was considered an unnecessary instruction that potentially impacted ratings. The procedure for the visual condition was similar, however, as described above, the first target photo was presented above the lineup and ratings made on the basis of perceptual similarity between the first target photo and the eight lineup photos. Participants in both the visual and memory ratings conditions involving the simultaneous lineup were instructed to:

*Please rate how much each person resembles the target person by entering a number from 1 to 7 on the following scale into each of the boxes underneath the corresponding photographs.*

The 7-point Likert-scale was labeled at three points, with ‘1’ equating to “Very little resemblance”, ‘4’ equating to “Moderate resemblance”, and ‘7’ equating to “Very high resemblance”. The sequential lineup procedure across all conditions involved photographs of lineup members being shown one at a time, with the following instruction:

*Please rate how much this person resembles the target person by entering a number from 1 to 7 on the following scale.*

When participants had completed a rating of one lineup member in the sequential lineups, the next lineup photo was presented. Ratings were made in relation to all eight lineup members in both SIML and SEQL procedures. Thus, the use of ratings in the SEQL formed a hybrid procedure that drew on components of the SIML and SEQL identification procedures. For instance, in contrast to the standard SEQL procedure used for identifications, participants completing ratings in the SEQL procedure viewed all lineup photos (which was necessary in order to complete ratings), and were made aware of how many photos they would view.
Results

Three measures of similarity are reported. Similar to study 1, the primary measure was the proportion in which the target or target replacement was rated most similar (received the highest rating) to a participant’s memory of the target. This involved transforming ratings data into binary data regarding whether the target was rated most similar (yes = 1) or not (no = 0). Whether the target was rated most similar or not was summed across the four target present lineups, providing a score between 0-4 for each participant. This process was also followed with regards to whether the target replacement was rated most similar across the four target absent lineups, providing a similar proportion target replacement score between 0-4 for each participant. Thus the main analyses were completed on mean proportion scores of between 0-4. In Table 8.2 these scores were transformed into a percentage figure under the ‘Highest Rating%’ – which simply involved dividing the mean proportion score by 4 to obtain a percentage. The ‘Highest Rating%’ are in bold as they were the main statistic that was compared to identification rates. Table 8.2 shows the mean proportions in percentage form (‘Highest Rating%’) for whether the target (lineups 1, 2, 5, and 6) or target replacement (lineups 3, 4, 7 and 8) were rated most similar. This information is provided for each individual lineup, overall target present, overall target absent, and according to early or late presentation of the target/target replacement. For instance, the target present mean column (TP Mean) shows the percentage in which targets received the highest similarity rating (Highest Rating%) for the Visual Task SIML (84.83%), Visual Task SEQL (96.75%), Memory Task SIML (53.58%), Memory Task SEQL (66.3%), as well as the equivalent conditions for the identification conditions (bottom rows). The second measure of similarity was employed in study 1 and entailed the average ratings of targets and target replacements. These are also included in Table 8.2 as “Average Rating” which is a figure based on the 0-7 point Likert-scale. This measure was used only infrequently due to the limitations outlined in study 1. The third measure of similarity was not included in study 1, and constituted the discrepancy between an individual rater’s highest and second highest scores, which allowed a determination of whether higher discrepancies were associated with greater accuracy (the target being rated most similar). Analyses entailed the comparison of mean proportions with univariate and repeated measures ANOVAs and t-tests using a probability level of $p < .05$. The assumption of homogeneity of variance was not violated unless specified. Non-parametric equivalents were also used to confirm significant results given that dependent variables regularly failed the assumption of normality.
Several other measures of similarity are reported briefly in Table 8.2 in order to provide further information, however, no statistical analyses were completed on these. The ‘Equal Highest Rating%’ indicates the percentage in which a lineup member was rated equally as highly as the target. The ‘Outrated%’ indicates the percentage in which another lineup member received a higher rating than the target/target replacement. Finally, the ‘All 1 Responses’ indicate the percentage of cases in which participants provided a similarity rating of 1/7 for every member of the lineup. Table 8.2 also includes the equivalent identification data from study 4 – including both studies in the one table allows for an easier visual comparison between rating and identification data.

General analyses were conducted on similarity ratings including the impact of: (1) visual versus memory task, (2) SIML/SEQL procedure, (3) the presence/absence of the target, (4) the order of presentation, (5) participant gender, (6) participant ethnicity, and (7) the discrepancy between an individuals highest and second highest rating. Analyses conducted in study 3 differed slightly to those in study 1 due to differences in the study designs.
### Similarity Ratings (Study 3)

<table>
<thead>
<tr>
<th></th>
<th>L1 Mean</th>
<th>L2 Mean</th>
<th>L3 Mean</th>
<th>L4 Mean</th>
<th>L5 Mean</th>
<th>L6 Mean</th>
<th>L7 Mean</th>
<th>L8 Mean</th>
<th>Mean Mean</th>
<th>Mean Early</th>
<th>Mean Late</th>
<th>TA Early</th>
<th>TA Late</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Average Rating</strong></td>
<td>6.52</td>
<td>6.17</td>
<td>3.78</td>
<td>3.39</td>
<td>6.57</td>
<td>5.87</td>
<td>3.17</td>
<td>2.96</td>
<td>6.29</td>
<td>3.33</td>
<td>6.55</td>
<td>6.02</td>
<td>3.48</td>
</tr>
<tr>
<td><strong>Highest Rating%</strong></td>
<td>95.7</td>
<td>87</td>
<td>21.7</td>
<td>21.7</td>
<td>87</td>
<td>69.6</td>
<td>21.7</td>
<td>21.7</td>
<td>84.83</td>
<td>21.7</td>
<td>91.35</td>
<td>78.3</td>
<td>21.7</td>
</tr>
<tr>
<td><strong>Equal Highest Rating%</strong></td>
<td>0</td>
<td>4.3</td>
<td>17.4</td>
<td>8.7</td>
<td>8.7</td>
<td>17.4</td>
<td>4.3</td>
<td>4.3</td>
<td>5.43</td>
<td>11.95</td>
<td>4.35</td>
<td>6.5</td>
<td>17.4</td>
</tr>
<tr>
<td><strong>Outrated%</strong></td>
<td>4.3</td>
<td>8.7</td>
<td>52.2</td>
<td>65.2</td>
<td>4.3</td>
<td>17.4</td>
<td>52.2</td>
<td>65.2</td>
<td>8.68</td>
<td>58.7</td>
<td>4.3</td>
<td>13.05</td>
<td>52.2</td>
</tr>
<tr>
<td><strong>All ‘1’ Responses%</strong></td>
<td>0</td>
<td>0</td>
<td>8.7</td>
<td>4.3</td>
<td>0</td>
<td>4.3</td>
<td>8.7</td>
<td>8.7</td>
<td>1.08</td>
<td>7.66</td>
<td>0</td>
<td>2.15</td>
<td>8.7</td>
</tr>
</tbody>
</table>

### Identifications (Study 4)

<table>
<thead>
<tr>
<th></th>
<th>L1 TP</th>
<th>L2 TP</th>
<th>L3 TA</th>
<th>L4 TA</th>
<th>L5 TP</th>
<th>L6 TA</th>
<th>L7 TA</th>
<th>L8 TA</th>
<th>Mean Mean</th>
<th>Mean Early</th>
<th>Mean Late</th>
<th>TA Early</th>
<th>TA Late</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Targets/TR identified%</strong></td>
<td>66.7</td>
<td>70.4</td>
<td>22.2</td>
<td>29.6</td>
<td>66.7</td>
<td>77.8</td>
<td>11.1</td>
<td>3.7</td>
<td>70.4</td>
<td>16.65</td>
<td>66.7</td>
<td>74.1</td>
<td>16.65</td>
</tr>
<tr>
<td><strong>Foils identified%</strong></td>
<td>14.8</td>
<td>7.4</td>
<td>14.8</td>
<td>33.3</td>
<td>11.1</td>
<td>18.5</td>
<td>25.9</td>
<td>29.6</td>
<td>12.95</td>
<td>25.9</td>
<td>12.95</td>
<td>20.35</td>
<td>31.45</td>
</tr>
<tr>
<td><strong>Misses%</strong></td>
<td>18.5</td>
<td>22.2</td>
<td>22.2</td>
<td>3.7</td>
<td>16.65</td>
<td>20.35</td>
<td>12.95</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Correct Rejections%</strong></td>
<td>63.0</td>
<td>37</td>
<td>63</td>
<td>66.7</td>
<td>57.43</td>
<td>63</td>
<td>51.85</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Memory Task SIML (n=27)

<p>| | | | | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Targets/TR identified%</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Foils identified%</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Misses%</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Correct Rejections%</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Memory Task SEQL (n=27)

<p>| | | | | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Targets/TR identified%</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Foils identified%</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Misses%</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Correct Rejections%</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
(1) The impact of condition: Visual versus memory task

Target Present.

It was expected that participants completing the visual perceptual task would rate the target most similar more often than participants completing the memory task. A 2 (visual/memory condition) x 2 (SIML/SEQL procedure) x 2 (participant gender) x 2 (participant ethnicity) univariate ANOVA was conducted, with mean proportion scores for whether the target was rated most similar (0-4) as the dependent variable.6 This ANOVA constituted the main analysis (alongside a replica ANOVA with mean proportion scores for whether target replacements were rated most similar), and is regularly referred to below, although other separate analyses were conducted to investigate specific manipulations of interest. The main effect of condition was highly significant, with participants who completed the visual task rating the target most similar more often (M = 3.532, SE = .162) than participants who completed the memory task (M = 2.374, SE = .152): F(1) = 27.214, p = <.001, d = 1.11, 95% CI of the difference [.716 – 1.600]. Levene’s test of equality of variances was violated. This finding was confirmed by a non-parametric independent samples Mann-Whitney U Test [U = 395.000, z = -5.331, p = <.001, r = .80]. No interactions involving condition were significant.

To further examine the impact of condition according to lineup procedure, independent samples t-tests were conducted. With regards to the SIML the main effect of the visual versus memory task was highly significant. Participants who completed the SIML visual task rated the target most similar more often (M = 3.39, SD = .839) than participants who completed the SIML memory task (M = 2.14, SD = 1.108); t(42) = 4.236, p < .001, 95% CI of the difference [.654 – 1.843]. Cohen’s effect size value (d = 1.27) indicated a large effect. This finding was confirmed by a non-parametric independent samples Mann-Whitney U Test [U = 89.500, z = -3.719, p = <.001, r = .56].

In regards to the SEQL the main effect of condition was also highly significant. Participants who completed the SEQL visual task rated the target most similar more often (M = 3.87, SD = .458) than participants who completed the SEQL memory task (M = 2.65, SD = 1.191); t(44) = 4.575, p < .001, 95% CI of the difference [.673 – 1.762]. Cohen’s effect size value indicated a large effect (d = 1.35). Levene’s test for equality of variances was violated for the SEQL. This finding

---

6 This analysis was first run with order included as a within participant variable, however, given no main effects or interactions involving order were significant, it was removed.
was confirmed by a non-parametric independent samples Mann-Whitney U Test \( U = 99.500, z = -4.138, p = <.001, r = .61 \).

**Target Absent.**

While the impact of the visual versus memory task was highly significant in regards to ratings of targets, this was not the case for ratings of target replacements. The same 2x2x2x2 univariate analysis used for the target present condition was repeated with the dependent variable changed to the mean proportion scores for whether the target replacement was rated most similar. Levene’s test of equality of variances was violated. No main effects or interactions involving condition approached significance. Target replacements were rated most similar almost equally often across the visual \( (M = .733, SE = .151) \) and memory \( (M = .756, SE = .142) \) tasks: \( F(1) = .013, p = .911 \).

**(2) The impact of lineup procedure: Simultaneous and sequential**

**Target present.**

Given the exploratory nature of this study no predictions were made in regards to the impact of lineup procedure on similarity ratings. The main 2x2x2x2 univariate ANOVA completed on ratings of targets indicated that the main effect of lineup procedure approached significance, with the target being rated most similar more often in the SEQL \( (M = 3.169, SE = .159) \) compared to SIML \( (M = 2.738, SE = .159) \): \( F(1) = 3.772, p = .056, d = .41 \). Levene’s test of equality of variances was violated. No interactions involving lineup procedure approached significance. When participant ethnicity and participant gender were removed from the analysis, the main effect for lineup procedure became significant. Targets in the SEQL were rated most similar significantly more often \( (M = 3.261, SE = .138) \) than targets in the SIML \( (M = 2.767, SE = .142) \): \( F(1) = 6.208, p = .015, d = .53, 95\% CI of the difference [.100 – .888] \). Levene’s test was not violated. This finding was confirmed by a non-parametric independent samples Mann-Whitney U Test confirmed \( [U = 1,269.000, z = 2.225, p = .026, r = .33] \).

To explore the above finding in more detail, independent samples t-tests were used to compare whether the target was rated most similar more often in the SIML or SEQL according to condition. For the visual similarity task the main effect of procedure was significant, with the target being rated most similar more often in the SEQL \( (M = 3.87, SD = .458) \) than the SIML \( (M = 3.39, SD = .839) \): \( t(44) = -2.400, p = .022, 95\% CI of the difference [-.883 – -.073], d = .71 \). The assumption of homogeneity of variance was violated. The significant result was confirmed by a
Identity Crisis in Identification Evidence

non-parametric independent samples Mann-Whitney U Test \([U = 355.000, z = 2.586, p = .010, r = .38]\). The above pattern was replicated in regards to the memory task, albeit in non-significant fashion. For the memory similarity task there was a tendency for the target to be rated most similar in the SEQL \((M = 2.65, SD = 1.191)\) compared to the SIML \((M = 2.14, SD = 1.108)\), though the difference was not significant; \(t(42) = -1.464, p = .151\).

The ‘Equal Highest Rating%’ measure presented in Table 8.2 shows that another lineup member was rated equally as similar as the target twice as often in the SIML as the SEQL. This was the case across both the visual and memory tasks. Targets were also outrated more often in the SIML compared to SEQL in the visual similarity task. This may indicate that participants had more difficulty perceiving the similarity of the target in the SIML compared to the SEQL.

**Target Absent.**

The main 2x2x2x2 univariate ANOVA completed on ratings of target replacements indicated that the main effect of lineup procedure was non-significant: \(F(1) = .034, p = .853\). The only significant result related to the interaction between procedure and participant gender: \(F(1) = 5.008, p = .028, d = .48\). This involved males rating target replacements most similar less often in the SIML compared to SEQL \((SIML, M = .550, SE = .251; SEQL, M = 1.052, SE = .233)\), with females showing the opposite trend \((SIML, M = .901, SE = .159; SEQL, M = .648, SE = .183)\).

(3) The impact of target presence/absence

Given that lineups 1, 2, 5 and 6 included targets and lineups 3, 4, 7, and 8 did not, it was not possible to compare the same lineup in both target present and absent conditions in this analysis as was done in study 1. A repeated measures ANOVA with the same 2x2x2x2 independent variable structure as above was conducted, with both the mean proportion scores for targets and target replacements included as dependent variables simultaneously. The main effect of target presence/absence was highly significant, with targets rated most similar clearly more often \((M = 2.953, SE = .111)\) than target replacements \((M = .745, SE = .104)\): \(F(1) = 215.410, p = <.001, d = 2.2, 95\% CI of the difference [1.909 – 2.508]\). Both dependent variables violated Levene’s test of equality or error variances. As shown in Table 8.2, targets were typically rated most similar at three to four times the rate of the target replacement across lineup procedure and condition. In the SIML visual task targets were rated most similar (84.83%) much more often
compared to the target replacement (21.7%). This was also the case for the SEQL visual task (target = 96.75%, target replacement = 13%), the SIML memory task (target = 53.58%, target replacement = 17.85%), and the SEQL memory task (target = 66.3%, target replacement = 19.58%).

Average ratings were also examined as they served to highlight how highly targets were rated compared to target replacements. One issue with comparing target present and absent data in the literature is that typically only target replacements are included as false positives in measures of accuracy, yet another lineup member (e.g. foil) may have been identified more often than the target replacement. This can present a misleading picture of accuracy when correct identifications are compared to false positives of the target replacement only. Applying this to ratings, one way around this was to treat the person with the highest average rating as the target replacement in target absent lineups in terms of analyses. This was done in this particular analysis, which compared the mean target rating with whomever had the highest average rating in the target absent condition (e.g. foil or target replacement). This therefore presented a more conservative estimate of the difference between target present and target absent ratings. The person rated on average most similar in target present lineups was always the target, and in target absent lineups was the target replacement in lineups 3, 4 and 7, and person 1 in lineup 8. These four averages were combined and compared across target present and absent conditions. Paired samples t-tests were conducted with average ratings as the dependent variable, which are shown in Table 8.3.

Table 8.3
*T-test results for target present and target absent average ratings according to procedure and condition*

<table>
<thead>
<tr>
<th>Average Ratings</th>
<th>M</th>
<th>SD</th>
<th>t</th>
<th>df</th>
<th>p</th>
<th>CI of the difference</th>
<th>d</th>
<th>r</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SIML Visual</strong></td>
<td>Target rating</td>
<td>6.28</td>
<td>1.23</td>
<td>11.89</td>
<td>22</td>
<td>&lt;.001</td>
<td>2.37 – 3.37</td>
<td>2.04</td>
</tr>
<tr>
<td></td>
<td>TA Highest rating</td>
<td>3.41</td>
<td>1.56</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>SIML Memory</strong></td>
<td>Target rating</td>
<td>4.79</td>
<td>1.75</td>
<td>4.09</td>
<td>20</td>
<td>.001</td>
<td>.74 – 2.45</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>TA Highest rating</td>
<td>3.17</td>
<td>1.47</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>SEQL Visual</strong></td>
<td>Target rating</td>
<td>6.45</td>
<td>.63</td>
<td>14.66</td>
<td>22</td>
<td>&lt;.001</td>
<td>3.15 – 4.18</td>
<td>4.12</td>
</tr>
<tr>
<td></td>
<td>TA Highest rating</td>
<td>2.78</td>
<td>1.08</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>SEQL Memory</strong></td>
<td>Target rating</td>
<td>5.24</td>
<td>1.34</td>
<td>4.03</td>
<td>22</td>
<td>.001</td>
<td>.74 – 2.31</td>
<td>1.16</td>
</tr>
<tr>
<td></td>
<td>TA Highest rating</td>
<td>3.72</td>
<td>1.30</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note:* For Average Ratings the target rating represents the average of the four target ratings. The TA Highest rating represents the average of the four persons rated most similar in the four target absent lineups.
The differences in average ratings between target present and absent conditions were highly significant across all four conditions. As evident in Table 8.3, the difference between target present and absent ratings was greatest for the visual (compared to memory) condition, and for the SEQL (compared to SIML). This effect was most pronounced for the SEQL visual condition, suggesting that participants both perceived targets to be more similar, and target replacement/foils in target absent lineups to be less similar compared to the other conditions.

Another measure of the difference between ratings in the target present and target absent lineups was the proportion of participants who provided all ‘1’ responses (‘Not at all similar”) for every member of a lineup. As shown in Table 8.2, there were more ‘All 1 responses’ in the target absent condition for the SIML visual task, SIML memory task, and SEQL visual task.

(4) The impact of order of presentation: Early versus late

Target Present.

If similarity ratings mimic identifications then it was anticipated that order effects would be more likely to occur in the SEQL procedure. Variables were collapsed to allow for a comparison of target present early and target present late scores. This provided a mean proportion score of between 0-2 for early presentation and 0-2 for late presentation for each participant (each participant completed four target present lineups, two early and two late). A repeated-measures ANOVA with the same 2x2x2x2 independent variable structure as above was conducted, but with both the mean proportion scores for targets presented early and targets presented late included as dependent variables simultaneously. There was no main effect for order. Targets presented early were rated most similar slightly more often ($M = 1.531, SE = .066$) than targets presented late ($M = 1.422, SE = .070$), though the difference was non-significant: $F(1) = 1.918, p = .170$. Levene’s test of equality or error variances was violated for both dependent variables. Whilst there was no main effect, the interaction between order and lineup procedure was highly significant: $F(1) = 12.170, p = .001, d = .52$. This was driven by participants in the SIML being more likely to rate targets presented early as most similar ($M = 1.561, SE = .095$) compared to targets presented late ($M = 1.117, SE = .100$). In comparison, targets in the SEQL were rated most similar approximately equally early ($M = 1.501, SE = .092$) and late ($M = 1.667, SE = .098$).

To further explore this interaction, paired samples t-tests were conducted separately on the SIML visual task, SIML memory task, SEQL visual task and SEQL
memory task. In the SIML visual task there was a main effect for order, with targets presented early ($M = 1.83, SD = .388$) being given the highest ratings more often than targets presented late ($M = 1.57, SD = .590$): $t(22) = 2.313, p = .03, 95\% CI of the difference [.027 – .495], d = .52, r = .25$. The significant result was confirmed by a non-parametric related samples Wilcoxon Signed Rank Test [$W = 4.500, z = 2.121, p = .034, r = .44$]. In the SEQL visual task there was a tendency for targets presented late ($M = 2.00, SD = .00$) to be rated most similar slightly more often than targets presented early ($M = 1.87, SD = .458$), but this difference was not significant: $t(22) = -1.367, p = .186$. In the SIML memory task there was no main effect for order, with targets presented early ($M = 1.14, SD = .655$) and late ($M = 1.00, SD = .837$) being rated most similar at roughly equivalent rates: $t(20) = .645, p = .526$. In the SEQL memory task there was again no effect for order, with targets presented early ($M = 1.30, SD = .635$) and late ($M = 1.35, SD = .714$) rated most similar at equivalent rates: $t(22) = -.327, p = .747$.

A separate frequency analysis was completed to determine whether the order in which participants completed the lineups impacted on their ratings. For this analysis the first and last target present lineups completed by each participant were compared. The dependent variable was whether the target had been rated most similar. In Table 8.4 the visual and memory conditions were collapsed to provide an indication of how many participants who completed the SIML ($N = 44$) and SEQL ($N = 46$) procedures rated the target most similar in the first compared to last target present lineup they completed. For instance, with regards to the SIML, 35 out of 44 participants rated the target most similar in the first target present lineup they completed, which dropped to 25 out of 44 for the last target present lineup they completed. As evident from Table 8.4, participants were around 20\% more likely to rate the target most similar in the first lineup they completed compared to the last. This was the case across both procedures, though the effect was stronger in the SIML procedure. This was surprising as it was expected that the SEQL would require more effort, and thus be more susceptible to the effects of fatigue. Given that participants completed the lineups in randomised order, these order effects are distributed evenly across conditions. Whilst such order effects impact the design and results of research studies (e.g. randomising the order in which lineups are presented is imperative), they are of less concern to the ratings process in practice, where typically a person will observe only one lineup.
Table 8.4

*Frequencies and proportions for whether the target was rated most similar in the first compared to last lineup completed by participants*

<table>
<thead>
<tr>
<th></th>
<th>First</th>
<th>Last</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SIML</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Frequency</td>
<td>35/44</td>
<td>25/44</td>
<td>10/44</td>
</tr>
<tr>
<td>Proportion</td>
<td>79.5</td>
<td>56.8</td>
<td>22.7</td>
</tr>
<tr>
<td><strong>SEQL</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Frequency</td>
<td>41/46</td>
<td>33/46</td>
<td>8/46</td>
</tr>
<tr>
<td>Proportion</td>
<td>89.1</td>
<td>71.7</td>
<td>17.4</td>
</tr>
</tbody>
</table>

*Target Absent.*

An equivalent repeated-measures ANOVA with the same 2x2x2x2 independent variable structure as above was conducted on target replacement rating data. Both the mean proportion scores for target replacements presented early and late were included as dependent variables simultaneously. The main effect of order was non-significant: \(F(1) = .129, p = .721\). No interactions involving order were significant.

(5) **The impact of participant gender**

*Target present.*

While there is some evidence to suggest increased ability to accurately identify persons of one’s own gender, this is inconclusive. As such, no predictions were made in relation to the impact of participant gender upon ratings. The main 2x2x2x2 univariate ANOVA completed with the mean proportion of targets rated most similar indicated that there was no main effect involving gender: \(F(1) = .921, p = .340\). However, when only participant gender and lineup procedure were included as independent variables, there was a significant interaction between the two: \(F(1) = 4.066, p = .047, d = .45\). Female participants rating SEQL targets most highly more often than males drove this interaction. Box’s test of equality of covariance matrices was violated, as was Levene’s test of equality or error variances. No other significant interactions involving participant gender were observed.

To further explore this, independent samples t-tests were conducted on whether the target was rated most similar using mean proportion scores (0-4). There were insufficient numbers of male participants to explore gender effects according to condition and procedure concurrently, so further t-tests were conducted that collapsed across condition but not lineup procedure. In relation to the SIML (visual and memory tasks combined), male participants tended to rate targets most similar slightly more often \((M = 2.93, SD = .917)\) than females \((M = 2.73, SD = 1.258)\).
though this difference did not approach significance: \( t(42) = .519, p = .607 \). In relation to the SEQL (visual and memory tasks combined), females rated targets most similar \( (M = 3.55, SD = .910) \) significantly more often than males \( (M = 2.76, SD = 1.200) \): \( t(44) = -2.514, p = .016, 95\% CI \text{ of the difference} [-1.418 \text{ to } -1.156], d = .60, r = .29 \). The significant result was confirmed by a non-parametric independent samples Mann-Whitney U Test \( U = 347.000, z = 2.611, p = .009, r = .39 \).

**Target Absent.**

The main 2x2x2x2 univariate ANOVA completed with the mean proportion of target replacements rated most similar indicated that there was no main effect involving gender: \( F(1) = .294, p = .590 \). However, similar to the target present condition, there was a significant interaction involving participant gender and lineup procedure: \( F(1) = 5.008, p = .028, d = .33 \). Independent-samples t-tests revealed that in relation to the SIML (visual and memory tasks combined), female participants tended to rate target replacements most similar \( (M = 3.196, SE = .131) \) significantly more often than males \( (M = 2.70, SE = .179) \): \( t(44) = 4.793, p = .032, d = .33 \). This finding was not supported by an independent-samples Mann-Whitney U Test \( U = 1,092.000, z = -2.812, p = .005, r = .42 \).

**(6) The impact of participant ethnicity**

**Target Present.**

The main 2x2x2x2 univariate ANOVA performed on mean proportions of whether the target was rated most similar indicated that there was a main effect for ethnicity. When all other variables were collapsed, Caucasian participants rated the target most similar significantly more often \( (M = 3.196, SE = .131) \) than non-Caucasian participants \( (M = 2.70, SE = .179) \): \( F(1) = 4.793, p = .032, d = .33 \). This finding was not supported by an independent-samples Mann-Whitney U Test \( U = 1,092.000, z = 1.613, p = .107 \), suggesting that the finding was not robust. No interactions involving participant ethnicity were significant. This finding, whilst not robust, was consistent with previous identification research that has demonstrated a bias towards own race, and suggests that similarity judgments for persons of another race may be more difficult. It was difficult to compare this finding to study 1 as study 1 included Asian lineups that were selected using a different process.
However, with this limitation in mind, these results were inconsistent with study 1, which showed a non-significant trend for Asian participants to rate Caucasian targets most highly slightly more often than Caucasian participants.

**Target Absent.**

The main 2x2x2x2 univariate ANOVA performed on mean proportions of whether the target replacement was rated most similar indicated that there was no main effect for ethnicity. Caucasian participants rated the target replacement as most similar ($M = .791, SE = .122$) slightly more often than non-Caucasian participants ($M = .699, SE = .168$), though the difference did not approach significance: $F(1) = .207, p = .659$. No interactions involving participant ethnicity were significant.

(7) **Discrepancy between highest and second highest ratings as an indicator of accuracy**

**Target Present.**

Drawing on Brewer et al.'s (2012) work, the discrepancy between an individual's highest and second highest scores were examined to determine whether higher discrepancies were associated with greater accuracy. Brewer found that greater discrepancy between the maximum and next highest ratings of participant's confidence that a photo was of the target was associated with correct decisions. A correct decision in relation to the current study was one in which the target was rated most similar. The relationship between discrepancy scores and accuracy was examined in relation to target present lineups. Discrepancy may prove a better indicator of accuracy than high ratings. For instance, consider two persons who rated the target most similar, one as 7/7, the other 4/7 in terms of similarity. Whilst it may seem that the person rating 7/7 provides more persuasive evidence, this interpretation may change in light of further information: e.g. that the first person's next highest rating was 6/7, where as the second persons next highest rating was 1/7.
Table 8.5

*The proportion of correct decisions based on the level of discrepancy between highest and second highest ratings for the memory task*

<table>
<thead>
<tr>
<th>Discrepancy</th>
<th>SEQL</th>
<th></th>
<th>SIML</th>
<th></th>
<th>SEQL &amp; SIML combined</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Proportion Correct</td>
<td>Number of decisions</td>
<td>Proportion Correct</td>
<td>Number of decisions</td>
<td>Proportion Correct</td>
<td>Number of decisions</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>1</td>
<td>100</td>
<td>5</td>
<td>83.3</td>
<td>6</td>
</tr>
<tr>
<td>5</td>
<td>100</td>
<td>5</td>
<td>100</td>
<td>3</td>
<td>100</td>
<td>8</td>
</tr>
<tr>
<td>4</td>
<td>100</td>
<td>7</td>
<td>100</td>
<td>5</td>
<td>100</td>
<td>12</td>
</tr>
<tr>
<td>3</td>
<td>93.4</td>
<td>16</td>
<td>87.5</td>
<td>8</td>
<td>91.7</td>
<td>24</td>
</tr>
<tr>
<td>2</td>
<td>87.5</td>
<td>24</td>
<td>71.4</td>
<td>14</td>
<td>81.6</td>
<td>38</td>
</tr>
<tr>
<td>1</td>
<td>42.9</td>
<td>28</td>
<td>53.6</td>
<td>28</td>
<td>48.2</td>
<td>56</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>11</td>
<td>0</td>
<td>21</td>
<td>0</td>
<td>32</td>
</tr>
</tbody>
</table>

*Note:* Discrepancy indicates the difference in a participant’s highest and second highest rating on the 7-point Likert-scale: e.g. a highest rating of 6/7 and next highest rating of 3/7 produces a discrepancy score of 3. A correct decision is one in which the target was rated most similar.

Table 8.5 shows the proportion of correct decisions based on the level of discrepancy between participant’s highest and second highest ratings for the memory task. Excluding discrepancies of six, of which one decision was incorrect in the SEQL, greater discrepancy scores were clearly associated with the target being rated most similar. Table 8.5 also shows that the SEQL procedure resulted in greater discrepancies compared to the SIML. For instance, in the SEQL 57.6% of decisions (or 53 out of 92) involved a discrepancy of two or more, of which 90% were correct. Whereas in the SIML 41.7% of decisions (or 35 out of 84) involved a discrepancy of two or more, of which 85.7% were correct. An independent samples t-test showed that whilst the SEQL memory task tended to result in a higher mean discrepancy ($M = 1.99, SD = 1.40$) compared to the SIML ($M = 1.73, SD = 1.70$), this difference was not significant: $t(174) = 1.125, p = .262$.

Table 8.6 shows the same data for the visual task. There were few errors in the visual task, rendering it less meaningful in terms of examining the proportion of correct responses. However, it was useful for demonstrating the increased number of higher discrepancies compared to the memory condition. For instance, in the SEQL visual task 80% of decisions (or 74 out of 92) involved a discrepancy of two or more, of which 100% were correct. In the SIML visual task 67.4% of decisions (or 62 out of 92) involved a discrepancy of two or more, of which 100% were correct. An independent samples t-test showed that the SEQL resulted in a significantly higher mean discrepancy ($M = 3.07, SD = 1.55$) compared to the SIML ($M = 2.46, SD = 1.58$) in the visual task: $t(174) = -2.549, p = .012$, 95% CI of the difference [-1.066 –
Identity Crisis in Identification Evidence

\[-.136\], \(d = .39\). The significant result was confirmed by a non-parametric independent samples Mann-Whitney U Test \([U = 4,733.000, z = 2.624, p = .009, r = .20]\).

Table 8.6
The proportion of correct decisions based on the level of discrepancy between highest and second highest ratings for the \textit{visual task}.

<table>
<thead>
<tr>
<th>Discrepancy</th>
<th>SEQL (Proportion Correct)</th>
<th>Number of decisions</th>
<th>SIML (Proportion Correct)</th>
<th>Number of decisions</th>
<th>SEQL &amp; SIML combined (Proportion Correct)</th>
<th>Number of decisions</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>100</td>
<td>5</td>
<td>100</td>
<td>7</td>
<td>100</td>
<td>12</td>
</tr>
<tr>
<td>5</td>
<td>100</td>
<td>17</td>
<td>100</td>
<td>2</td>
<td>100</td>
<td>19</td>
</tr>
<tr>
<td>4</td>
<td>100</td>
<td>11</td>
<td>100</td>
<td>9</td>
<td>100</td>
<td>20</td>
</tr>
<tr>
<td>3</td>
<td>100</td>
<td>24</td>
<td>100</td>
<td>20</td>
<td>100</td>
<td>44</td>
</tr>
<tr>
<td>2</td>
<td>100</td>
<td>17</td>
<td>100</td>
<td>24</td>
<td>100</td>
<td>41</td>
</tr>
<tr>
<td>1</td>
<td>82.4</td>
<td>17</td>
<td>65.2</td>
<td>23</td>
<td>72.5</td>
<td>40</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>7</td>
<td>0</td>
<td>8</td>
</tr>
</tbody>
</table>

An independent samples t-test was also conducted to determine the difference in mean discrepancy scores between the visual and memory tasks. The SIML and SEQL procedure were collapsed in this analysis. The visual task \((M = 2.72, SD = 1.59)\) resulted in a greater mean discrepancy score compared to the memory task \((M = 1.86, SD = 1.55)\), which was highly significant: \(t(358) = 5.187, p < .001, 95\% CI of the difference \([.533 – .1.185]\), \(d = .55\). The significant result was confirmed by a non-parametric independent samples Mann-Whitney U Test \([U = 111,047.000, z = 5.317, p < .001, r = .28]\).

\textit{Target Absent.}

In the target absent condition, an examination of discrepancies was conducted, but with two differences. First, in terms of ‘accuracy’ there was no correct decision equivalent as the target was not present. A target replacement receiving the highest rating could have been treated as a correct response, but this was essentially meaningless. It was a meaningless measure because while six researchers rated the target replacement as most resembling the target, target replacements were not as differentiated as targets. This meant that there was more scope for variability in regards to perceptions of similarity in target absent lineups. This was evident in the fact that target replacements overall were rated most similar between 13% and 21.7% across the four ratings conditions. Second, where in target present lineups discrepancy was between the target and the next highest rating, in the target absent lineups discrepancy was between any highest and second highest
Identity Crisis in Identification Evidence

rating. Table 8.7 depicts target absent discrepancies for the memory and visual tasks and SEQL and SIML procedures. It is clear that in target absent lineups, the majority of discrepancies are 0 and 1. In fact in 76.4% of target absent lineups the discrepancy between the highest and second highest rating is 0 or 1. Comparatively, in target present lineups, only 37.8% of discrepancies are 0 or 1. Also apparent is that there was again greater discrepancy with regards to the SEQL.

Table 8.7
Discrepancy frequencies for target absent lineups for the memory and visual tasks and SEQL and SIML procedure

<table>
<thead>
<tr>
<th>Discrepancy</th>
<th>Memory Task</th>
<th>Visual Task</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SEQL</td>
<td>SIML</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>14</td>
<td>9</td>
</tr>
<tr>
<td>1</td>
<td>45</td>
<td>30</td>
</tr>
<tr>
<td>0</td>
<td>24</td>
<td>38</td>
</tr>
</tbody>
</table>

Note: % Overall shows the percentage of discrepancy scores when all conditions were collapsed.

To explore differences in discrepancy scores across target present and absent lineups paired samples t-tests were conducted for each condition. As shown in Table 8.8, the difference in discrepancy scores was significant between target present and absent lineups across all four conditions. Lineups in which the target was present lead to larger discrepancy scores. This suggests that low discrepancy scores may be indicative of lineups in which a suspect is absent, or at least lineups in which the witness fails to perceive one person as being noticeably more similar.

Table 8.8
Paired samples t-tests comparing mean discrepancy scores in target present/absent lineups

<table>
<thead>
<tr>
<th>Target present</th>
<th>SEQL Memory</th>
<th>Mean</th>
<th>Std. Deviation</th>
<th>N</th>
<th>t</th>
<th>df</th>
<th>p</th>
<th>Cohen’s d</th>
</tr>
</thead>
<tbody>
<tr>
<td>Present</td>
<td>1.99</td>
<td>1.40</td>
<td>92</td>
<td>5.15</td>
<td>91</td>
<td>&lt; .001</td>
<td>0.71</td>
<td></td>
</tr>
<tr>
<td>Absent</td>
<td>1.13</td>
<td>1.01</td>
<td>92</td>
<td>4.50</td>
<td>83</td>
<td>&lt; .001</td>
<td>0.65</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Target present</th>
<th>SIML Memory</th>
<th>Present</th>
<th>1.73</th>
<th>1.70</th>
<th>84</th>
<th>4.50</th>
<th>83</th>
<th>&lt; .001</th>
<th>0.65</th>
</tr>
</thead>
<tbody>
<tr>
<td>Absent</td>
<td>0.83</td>
<td>0.97</td>
<td>84</td>
<td>1.24</td>
<td>1.47</td>
<td>92</td>
<td>5.57</td>
<td>&lt; .001</td>
<td>0.60</td>
</tr>
</tbody>
</table>
An important question resulting from the above analysis was that while discrepancies were clearly greater in target present lineups, did these greater discrepancies involve the target being rated most similar? To answer this question, univariate ANOVAs were conducted on the four target present conditions with discrepancy scores as the dependent variable, and whether the target was rated most similar (Yes/No) as the independent variable. As shown in Table 8.9, discrepancies were significantly higher in all four conditions when the target was rated most similar. This suggests that higher discrepancy scores pertained to the target being rated most similar, and provides further evidence that discrepancy scores can be used to index accuracy in regards to ratings. Also apparent from Table 8.9 is that the discrepancy scores for foils were higher in the SEQL compared to SIML. Thus the tendency towards larger discrepancies in the SEQL also pertained to situations where foils were rated most similar.

Table 8.9

<table>
<thead>
<tr>
<th>Rated most similar</th>
<th>Mean</th>
<th>Std. Deviation</th>
<th>N</th>
<th>F</th>
<th>df</th>
<th>p</th>
<th>r</th>
<th>Levene’s Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>SEQL Memory Target</td>
<td>2.53</td>
<td>1.19</td>
<td>60</td>
<td>36.52</td>
<td>1</td>
<td>&lt; .001</td>
<td>1.34</td>
<td>.06</td>
</tr>
<tr>
<td>Foil</td>
<td>0.97</td>
<td>1.18</td>
<td>32</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SIML Memory Target</td>
<td>2.69</td>
<td>1.70</td>
<td>45</td>
<td>48.82</td>
<td>1</td>
<td>&lt; .001</td>
<td>1.55</td>
<td>.00</td>
</tr>
<tr>
<td>Foil</td>
<td>0.62</td>
<td>0.78</td>
<td>39</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SEQL Visual Target</td>
<td>3.17</td>
<td>1.50</td>
<td>88</td>
<td>10.34</td>
<td>1</td>
<td>.002</td>
<td>1.66</td>
<td>.04</td>
</tr>
<tr>
<td>Foil</td>
<td>0.75</td>
<td>0.50</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SIML Visual Target</td>
<td>2.74</td>
<td>1.45</td>
<td>77</td>
<td>33.86</td>
<td>1</td>
<td>&lt; .001</td>
<td>1.66</td>
<td>.01</td>
</tr>
<tr>
<td>Foil</td>
<td>0.53</td>
<td>0.52</td>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: ‘Rated most similar’ indicates whether discrepancy scores pertain to ratings where the target was rated most similar or where a foil was rated most similar.

Finally, a univariate ANOVA was conducted to determine whether discrepancy scores regarding target present lineups differed across the four independent variables. This analysis utilised the same 2x2x2x2 independent variable structure, with the dependent variable being mean discrepancy scores for the target present lineups (e.g. each participants’ four discrepancy scores for the four target present lineups were averaged to produce one discrepancy score per participant). The only significant main effect was for condition, with discrepancy scores in the visual task being significantly higher ($M = 2.717$, $SE = .192$) than the memory task ($M = 1.866$, $SE = .180$): $F(1) = 10.421$, $p = .002$, $d = .48$. While there as a trend for the SEQL to produce higher discrepancy scores ($M = 2.497$, $SE = $...
.184) compared to the SIML ($M = 2.086, SE = .189$), this was not significant: $F(1) = 2.439, p = .123$. Females produced higher discrepancy scores ($M = 2.457, SE = .148$) than males ($M = 2.127, SE = .218$) on average, but this was not significant: $F(1) = 1.570, p = .214$. Caucasian ($M = 2.394, SE = .155$) and non-Caucasian ($M = 2.279, SE = .213$) participants produced equivalent discrepancy scores in target present lineups: $F(1) = .009, p = .925$. An equivalent univariate ANOVA conducted on target absent discrepancy scores indicated that no main effects approached significance.

CHAPTER 9.

Study 4

Method

Study 4 involved the traditional identification procedure for the SIML and SEQL lineup procedures. The main purpose of study 4 was to provide a point of comparison, or reference point, for study 3. Thus, study 4 replicated study 3 in many ways, and these two studies formed part of a larger research project. The comparison of study 3 and study 4 data occurs in Chapter 10 below. A secondary aim of study 4 was to explore the impact of the independent variables on identification data (hits, misses, false positives and correct rejections). As such, this chapter describes the study 4 method and reports exclusively on identification data.

Participants

Across both SIML and SEQL identification conditions, 74 persons accessed the survey page, of which 73% completed the study. The majority of dropouts occurred prior to the completion of lineup 1. The discrepancy between completion rates for study 3 and study 4 mirrored that found between study 1 and study 2, and reasons for the discrepancy are discussed there. The final sample therefore comprised a total of 54 participants who completed the identification task in study 4, including 27 participants in each lineup procedure. They were aged 17-64 years ($M = 31.5, SD = 13.3$), and were 72.2% female. Caucasian participants constituted 64.8% of the sample. Recruitment processes mirrored those of study 1.

Design

The study 4 design replicated that of study 3, but rather than making resemblance ratings participants either selected one person from the lineup, or indicated that the target was ‘not present’. Study 4 did not include the visual
condition from study 3, meaning that participants only made memory-based judgments. A visual identification condition was not included as it was expected that it would lead to a near 100% correct identification rate, making it redundant. Thus, study 4 incorporated a 2 (lineup procedure: SIML/SEQL) x 2 (target: present or absent) x 2 (order: early/late presentation) x 2 (participant gender) x 2 (participant ethnicity) mixed design, which is outlined in Table 9.1. Lineup procedure, participant gender and participant ethnicity were between participant variables, with all remaining being within.

Table 9.1

Study 4 Experimental Design

<table>
<thead>
<tr>
<th>Identity</th>
<th>SIML (n=27)</th>
<th>Memory Task</th>
<th>Lineup 1</th>
<th>Lineup 2</th>
<th>Lineup 3</th>
<th>Lineup 4</th>
<th>Lineup 5</th>
<th>Lineup 6</th>
<th>Lineup 7</th>
<th>Lineup 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Memory</td>
<td>Present / Early</td>
<td>Present / Late</td>
<td>Absent / Early</td>
<td>Absent / Late</td>
<td>Present / Early</td>
<td>Present / Late</td>
<td>Absent / Early</td>
<td>Absent / Late</td>
<td>Absent / Early</td>
<td>Absent / Late</td>
</tr>
</tbody>
</table>

Note: All lineups exclusively included photos of Caucasian males. SIML indicates the simultaneous lineup, and SEQL indicates the sequential lineup procedure.

Materials

The materials in study 4 replicated those in study 3.

Procedure

The procedure in study 4 replicated the memory conditions of study 3, but instead of ratings participants made traditional identification decisions, or indicated that the target was not present. As noted above, there was also no visual condition. In the SEQL procedure in study 3, ratings were given to every lineup member. However, in study 4, the lineup ended when an identification was made, which is consistent with the traditional SEQL procedure. This meant that not all SEQL lineup photos were necessarily seen. Similarly, where participants in study 3 were explicitly advised that they would view eight lineup photos, participants completing identifications in study 4 did not receive this instruction. This was because ‘backloading’ (participant’s being unaware of how many photos they will view) has been suggested to be a principal component of the SEQL identification procedure (Zimmerman et al., 2006).

Results

Dependent variables constituted hits, false positives and misses in the target present condition, and correct rejections in the target absent condition. As in study
2, any identification in a target absent lineup was treated as a false positive. Study 2 above reported target absent false positives, however, study 4 reports correct rejections (each being the inverse of the other). Similar to study 2, dependent variables were a score between 0-4 for analyses not including order of presentation, and a score between 0-2 for analyses including order (early versus late). For example, for the hits dependent variable, a participant’s four target present responses were coded as either a hit (coded ‘1’) or non-hit (coded ‘0’), which were then summed (providing a score between 0-4). This same process was followed in relation to false positives and misses in target present lineups, and correct rejections in target absent lineups. Analyses were completed on each dependent variable measure separately. Thus four univariate ANOVAS were completed (e.g. one each for hits, target present false positives, misses, and correct rejections). Four repeated measures ANOVAs were similarly conducted when early versus late presentation was being examined. Significant results were followed up with independent-samples t-tests where appropriate, using a probability level of \( p < 0.05 \). Non-parametric equivalents were also used to confirm significant results. Cell sizes were \( \geq 10 \) with the following exceptions: only four males completed the SIML procedure, and only six non-Caucasian participants completed the SEQL. As such, analyses involving participant gender and ethnicity were difficult, and results involving them were interpreted with caution. Table 8.2 above shows the breakdown of hits, misses, false positives, and correct rejections according to lineup. Results addressed the following questions: (1) the impact of lineup procedure, (2) the impact of order of presentation, (3) the impact of gender, and (4) the impact of ethnicity.

(1) The impact of lineup procedure

**Target Present Hits.**

A 2 (SIML/SEQL) x 2 (participant gender) x 2 (participant ethnicity) univariate ANOVA was conducted with target present hits (between 0-4) as the dependent variable.\(^7\) There was no main effect for lineup procedure in relation to hits, with hits being roughly equivalent between the SIML (\( M = 2.826, SE = .357 \)) and SEQL (\( M = 2.945, SE = .342 \)): \( F(1) = .059, p = .810 \). Levene’s test of equality of error variances was violated. No other main effects or interactions were significant.

---

\(^7\) This analysis was first run with order included as a within participant variable, however, given no main effects or interactions involving order were significant, it was removed. The impact of order is explored in section (2).
Target Present False Positives.

The above analysis was repeated with target present false positives as the dependent variable (providing a score between 0-4). There was no main effect for lineup procedure, with slightly fewer false positives occurring in SIML ($M = .403, SE = .225$) compared to SEQL lineups ($M = .443, SE = .217$): $F(1) = .016, p = .899$. No other main effects or interactions were significant.

Target Present Misses.

The above analysis was repeated with misses in target present lineups as the dependent variable (providing a score between 0-4). While there was a trend for fewer misses to occur in the SIML ($M = .456, SE = .260$) compared to SEQL ($M = .611, SE = .249$), this was not significant: $F(1) = .185, p = .669$. No other main effects or interactions were significant.

Target Absent Correct Rejections (and False Positives).

The above analysis was repeated with correct rejections in target absent lineups as the dependent variable (providing a score between 0-4). While fewer correct rejections occurred in the SIML ($M = 1.659, SE = .342$) compared to SEQL ($M = 2.077, SE = .329$), this difference was not significant: $F(1) = .776, p = .383$. However, the interaction between lineup procedure and participant gender was significant, with males making fewer correct rejections in the SIML ($M = .833, SE = .644$) compared to SEQL ($M = 2.700, SE = .585$), and females making more correct rejections in the SIML ($M = 2.485, SE = .233$) compared to SEQL ($M = 1.455, SE = .301$): $F(1) = 9.316, p = .004, d = .59$. However, there were only four males in the SIML, meaning that these results were interpreted with caution. Target absent false positive results inversely mirror those of the target absent correct rejections. As such, they are not repeated.

(2) The impact of order of presentation

Each participant completed two lineups in which the target was presented early, and two in which the target was presented late. These were summed for each participant, providing an early score between 0-2 and late score between 0-2. These two dependent variables were then compared using a repeated measures ANOVA. While in study 2 order was a between participant variable, in study 4 it formed a within participant variable. Participant gender and ethnicity were excluded from this analysis for two reasons. First, the question of interest was whether order effects existed as a function of lineup procedure (differences between the SIML and SEQL).
Thus, the focus was on the interaction between order and procedure. As a consequence, the main effect for order was not of interest in the repeated measures analysis, as it collapsed across the SIML and SEQL procedures. Instead it was explored using paired samples t-tests conducted on each lineup procedure separately. Second, the inclusion of participant gender and ethnicity reduced some cell sizes to < 5 (e.g. only one non-Caucasian male completed the SIML). The independent variables in this analysis were lineup procedure and early/late presentation.

**Target Present Hits.**

There was a significant interaction between order and procedure for hits: $F(1) = 5.145, p = .027, d = .63$. There was little difference in the SIML between early ($M = 1.333, SE = .127$) and late ($M = 1.370, SE = .140$) presentation. However, in the SEQL more targets were identified early ($M = 1.593, SE = .127$) compared to late ($M = 1.148, SE = .140$). Box’s test of equality of covariance matrices was violated. Two paired samples t-test were completed on each lineup procedure separately to check for main effects relating to order. The difference between early and late presentation in the SIML was not significant for hits: $t(26) = -.372, p = .713$. The difference between early and late presentation in the SEQL was significant for hits: $t(26) = 2.371, p = .025, d = .62$. A related samples Wilcoxon Signed Rank Test confirmed the difference as significant: $W = 37.500, z = -2.216, p = .027, r = .43$.

**Target Present False Positives.**

The same repeated measures ANOVA was completed with early and late target present false positives as the dependent variables. There was a greater difference between early and late false positives in the SEQL (Early, $M = .148, SE = .085$; Late $M = .333, SE = .104$) compared to SIML (Early, $M = .222, SE = .085$; Late $M = .259, SE = .104$), however, the interaction between order and procedure was not significant: $F(1) = .669, p = .417$. Two paired samples t-test were completed on each lineup procedure separately to check for main effects relating to order. The difference between early and late presentation was not significant for target present false positives with regards to the SEQL [$t(26) = -1.308, p = .202$], or SIML [$t(26) = -.328, p = .746$].

**Target Present Misses.**

The same repeated measures ANOVA was completed with early and late target present misses as the dependent variables. The interaction between order and procedure was significant: $F(1) = 4.382, p = .041, d = .58$. In the SIML more
misses occurred early \((M = .407, SE = .099)\) compared to late \((M = .259, SE = .113)\), whereas in the SEQL more misses occurred late \((M = .519, SE = .113)\) compared to early \((M = .259, SE = .099)\). Both dependent variables narrowly violated Levene’s test of equality of error variances. Two paired samples t-tests were completed on each lineup procedure separately to follow up the significant finding. The difference between early and late presentation was not significant for misses with regards to the SIML \([t(26) = 1.280, p = .212]\), or SEQL \([t(26) = -1.657, p = .110]\).

**Target Absent Correct Rejections (and False Positives).**

The same repeated measures ANOVA was completed with early and late target absent correct rejections. The interaction between order and procedure was not significant: \(F(1) = .027, p = .870\). There was a slight trend across both procedures for more lineups to be correctly rejected when the target replacement was presented early (SIML early, \(M = 1.259, SE = .146\), SIML late, \(M = 1.037, SE = .132\); SEQL early, \(M = 1.111, SE = .146\), SEQL late, \(M = .852, SE = .132\)). Two paired samples t-tests were completed on each lineup procedure separately to check whether this trend was significant. With regards to the SIML the difference between early and late presentation for correct rejections was not significant: \(t(26) = 1.237, p = .222\). With regards to the SEQL the difference between early and late presentation for correct rejections approached significance: \(t(26) = 1.892, p = .070\).

(3) The impact of participant gender

These results regarding the impact of participant gender were obtained from the initial univariate ANOVA reported in section (1) above. Whilst no main effects or interactions involving gender were significant (excluding the interaction between procedure and gender which is described above), results are reported here in more detail.

**Target Present Hits.**

As noted, only four males completed the SIML, meaning that these results were interpreted with caution. With this limitation in mind, the univariate ANOVA completed above on target present hits (0-4) indicated that whilst males tended to make more hits \((M = 3.100, SE = .453)\) than females \((M = 2.671, SE = .198)\), this difference was not significant: \(F(1) = .752, p = .390\). No interactions involving gender were significant. Levene’s test of equality of error variances was violated.
Target Present False Positives.

Males tended to make fewer false positives ($M = .292$, $SE = .286$) than females ($M = .555$, $SE = .125$), however, this difference was not significant: $F(1) = .709$, $p = .404$. No interactions involving gender were significant.

Target Present Misses.

Males tended to make fewer misses ($M = .358$, $SE = .330$) than females ($M = .709$, $SE = .144$), however, this difference was not significant: $F(1) = .951$, $p = .335$. No interactions involving gender were significant.

Target Absent Correct Rejections (and False Positives).

Males tended to make fewer correct rejections ($M = 1.767$, $SE = .435$) than females ($M = 1.970$, $SE = .190$), however, this difference was not significant: $F(1) = .183$, $p = .671$. As described above, the interaction between lineup procedure and participant gender was significant.

(4) The impact of participant ethnicity

These results regarding the impact of participant ethnicity were obtained from the initial univariate ANOVA reported in section (1) above. Results are reported here in more detail.

Target Present Hits.

Non-Caucasian participants tended to make slightly more hits ($M = 2.967$, $SE = .439$) than Caucasians ($M = 2.805$, $SE = .228$), though this difference did not approach significance: $F(1) = .108$, $p = .744$. Levene’s test of equality of error variances was violated. No interactions involving participant ethnicity were significant with regards to hits.

Target Present False Positives.

Caucasian participants tended to make slightly more target present false positives ($M = .451$, $SE = .144$) than non-Caucasians ($M = .396$, $SE = .277$), though this difference did not approach significance: $F(1) = .031$, $p = .861$. A trend was observed for Caucasian participants to make more false positives in the SIML ($M = .515$, $SE = .239$) compared to SEQL ($M = .386$, $SE = .160$), whereas non-Caucasian participants made more false positives in the SEQL ($M = .500$, $SE = .402$) compared to SIML ($M = .292$, $SE = .382$). But this interaction between lineup procedure and participant gender was not significant: $F(1) = .291$, $p = .592$. 
Target Present Misses.

There was no main effect for participant ethnicity with regards to misses: $F(1) = 1.091, p = .302$. Both Caucasian and non-Caucasian participants tended to make fewer misses in the SIML than SEQL but the interaction did not approach significance: $F(1) = .017, p = .898$.

Target Absent Correct Rejections (and False Positives).

Caucasian participants tended to make more correct rejections ($M = 2.153, SE = .219$) than non-Caucasians ($M = 1.583, SE = .421$), but the difference was not significant: $F(1) = 1.441, p = .236$. The interaction between procedure and participant ethnicity was not significant, but there was a trend for Caucasian participants to correctly reject the lineup equivalently across the two lineup procedures (SIML, $M = 2.152, SE = .363$; SEQL, $M = 2.155, SE = .244$), whereas non-Caucasian participants correctly rejected the lineup more often in the SEQL ($M = 2.000, SE = .611$) than SIML ($M = 1.167, SE = .580$): $F(1) = .765, p = .386$.

CHAPTER 10.

Comparing Study 3 & Study 4

Results (Studies 3 & 4)

A descriptive comparison of similarity (study 3) and identification (study 4) results is provided. Table 8.2 shows the proportions of similarity ratings and identifications for lineups 1-8 for the SIML and SEQL procedure and visual and memory conditions. Keeping in mind that there was no visual identification condition (identifications were memory-based only), the conditions were most accurate (correct identification or target rated most similar) in the following order: SEQL visual task, SIML visual task, SIML identification task, SEQL identification task, SEQL memory task, SIML memory task. When the memory-based ratings and identifications were compared, overall, targets were identified more often (69.5%) than they were rated most similar (59.9%). When the same comparison was made according to lineup procedure, SEQL ratings were more diagnostic of accuracy compared to the SIML. In the SEQL memory task the target was rated most similar 66.3% of the time and was identified from SEQL lineups 68.5% of the time. Comparatively, in the SIML memory task the target was rated most similar 53.6% of the time and was identified from SIML lineups 70.4% of the time. That a stronger relationship existed between SEQL memory task ratings and identifications
Identity Crisis in Identification Evidence

compared to SIML memory task ratings was unexpected. It was anticipated that SIML memory task ratings would better correspond to identifications, as not all photos are necessarily viewed in the SEQL identification condition, particularly when the target is presented late, making the comparison between SEQL ratings and identifications less complete.

A problem with the SEQL identification procedure: The target may not be seen

As noted above, one problem with the SEQL is that not all lineup members are necessarily seen as the lineup ends when an identification is made. This means that a foil may be identified prior to the target being observed. In the current study, foils appearing prior to the target were identified in 8.3% of cases. In the corresponding similarity condition, the SEQL memory task, a foil appearing prior to the target received the highest rating in 6.5% of cases. Interestingly, these figures increased in the target absent condition, where, in 24.1% of cases a foil was identified prior to the target replacement being observed. In the SEQL memory task, a foil that appeared prior to the target replacement received the highest rating in 25% of cases. Participants who made these identifications were unaware of whether the target was present or absent from the lineup, thus it is unclear why the rate increased in the target absent condition. That in 8.3% of target present lineups and 24.1% of target absent lineups, participants completing the SEQL identification procedure did not even observe the target/target replacement poses a real practice issue for the SEQL. Ways to reduce this problem are discussed later in the paper.

Differentiation and Match

As in the study 1 and 2 comparison, the concepts of differentiation and match were used to examine in more detail the relationship between ratings and identifications. Differentiation and match were operationalised in the same way as previously, and the same three things were of interest: overall levels of match (whether the person rated most similar was most identified), overall levels of differentiation (whether a person stood out), and how often differentiation predicted match (a more precise measure of the relationship between ratings and identifications than match alone).\(^8\) Match being high would evidence a stronger relationship between ratings and identifications, however, it was expected that match would be more likely to be high when a person was moderately or highly differentiated. For instance, if no one person was rated most similar most often then it was logically impossible for this nonexistent person to be identified most often.

\(^8\) See pages 60-62 above for definitions of high/moderate/low differentiation and match.
Visual and memory task.

Comparisons of match were made between each visual and memory task and the corresponding lineup procedure. Thus, SIML visual-task ratings were compared to SIML identifications, and SIML memory-task ratings were also compared to SIML identifications. The same was done for the SEQL. The level of differentiation and match were compared between the visual task and identifications and the memory task and identifications separately. These are depicted in Table 10.1. This allowed an examination of whether there was a stronger relationship between ratings and identifications in the visual or memory tasks. It produced 32 comparisons in total. It was predicted that if perceptual similarity was the basic building block upon which ecphoric judgments were made then the visual task ratings would better match identifications than the memory task ratings.

Table 10.1

<table>
<thead>
<tr>
<th>Level of Match</th>
<th>High-M</th>
<th>Mod-M</th>
<th>Low-M</th>
<th>Total</th>
<th>Predictability%</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Visual Task</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level of Differentiation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High-D</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>100</td>
</tr>
<tr>
<td>Mod-D</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Low-D</td>
<td>1</td>
<td>5</td>
<td>2</td>
<td>8</td>
<td>25</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>9</td>
<td>5</td>
<td>2</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td><strong>Memory Task</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level of Differentiation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High-D</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>100</td>
</tr>
<tr>
<td>Mod-D</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Low-D</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>7</td>
<td>14.3</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>11</td>
<td>4</td>
<td>1</td>
<td>16</td>
<td></td>
</tr>
</tbody>
</table>

Note: This table compares the SIML visual ratings task to the SIML identification task, and the SEQL visual ratings task to the SEQL identification task (16 comparisons), and the SIML memory ratings task to the SIML identification task, and the SEQL memory ratings task to the SEQL identification task (16 comparisons). The predictability column indicates how well the level of differentiation predicted the level of match.

Overall, differentiation was high in 16/32 cases, moderate in 1/32, and low in 15/32. There was little difference in levels of differentiation across both tasks: 8/16 persons were highly differentiated in both the visual and memory tasks. Overall, levels of match were high in 20/32 cases, moderate in 9/32, and low in 3/32. That in only three of 32 lineups the level of match between ratings and identifications was low provides strong evidence for a relationship between similarity ratings and identifications. Furthermore, this pattern is consistent with the results of studies 1 and 2. Counter to expectations, levels of match were slightly higher in the memory task (11/16) compared to the visual task (9/16), suggesting that there was a slightly
stronger association between the memory task and identifications compared to the visual task and identifications. Overall, differentiation predicted match in 19/32 cases, with a discrepancy of one category (e.g. Low-D and Mod-M) in 9/32, and a discrepancy of two categories (e.g. Low-D and High-M) in 3/32. Instances where differentiation predicted match are highlighted in Table 10.1. Perhaps the strongest evidence for a relationship between similarity ratings and identifications was that High-D predicted High-M 100% of the time. While overall match was slightly higher in relation to the memory task and identifications compared to the visual task, differentiation predicted match slightly more often in relation to the visual task (10/16) compared to the memory task (9/16). However, the small size of both differences suggests that the relationship between ratings and identifications was similar across both visual and memory tasks.

SIML and SEQL procedure.

The levels of differentiation and match were also compared between the SIML and SEQL procedures (Table 10.2). This allowed an examination of whether there was a stronger relationship between ratings and identifications in the SIML or SEQL. This again produced 32 comparisons in total. There was little difference between procedures. Differentiation was high in 8/16 cases across both procedures. Match was high in 10/16 for the SIML and 11/16 for the SEQL. Differentiation predicted match in 10/16 cases for the SIML, and 9/16 cases for the SEQL. High-D predicted High-M 100% of the time across both procedures.

Table 10.2

Frequencies for level of differentiation and match, and percentages for how well level of differentiation predicted level of match for the SIML and SEQL

<table>
<thead>
<tr>
<th>Level of Match</th>
<th>SIML</th>
<th>SEQL</th>
</tr>
</thead>
<tbody>
<tr>
<td>High-D</td>
<td>8 0 0</td>
<td>8 0 0</td>
</tr>
<tr>
<td>Mod-D</td>
<td>0 0 0</td>
<td>0 0 0</td>
</tr>
<tr>
<td>Low-D</td>
<td>2 4 2</td>
<td>1 5 1</td>
</tr>
<tr>
<td>Total</td>
<td>10 4 2</td>
<td>10 5 1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Level of Differentiation</th>
<th>Predictability%</th>
</tr>
</thead>
<tbody>
<tr>
<td>High-D</td>
<td>100</td>
</tr>
<tr>
<td>Mod-D</td>
<td>0</td>
</tr>
<tr>
<td>Low-D</td>
<td>25</td>
</tr>
<tr>
<td>Total</td>
<td>25</td>
</tr>
</tbody>
</table>

Note: This table compares the SIML visual and memory ratings task to the SIML identification task (16 comparisons) and the SEQL visual and memory ratings task to the SEQL identification task (16 comparisons). The predictability column indicates how well the level of differentiation predicted the level of match.
Target present compared to target absent.

As in study 1, the differentiation and match were examined according to the presence/absence of the target. In target present lineups, 16/16 were High-D. In comparison there were no High-D persons in target absent lineups. Match was high in 16/16 target present lineups, and 4/16 target absent lineups. These findings indicate that ratings were more diagnostic of identifications when the target was present. This likely reflects the fact that greater ecphoric similarity was perceived when the target was present, and is consistent with the hypothesis that ecphoric similarity underpins identifications.

Graphic representation of the 8 lineups

The eight lineups are graphically presented in Appendix C. This demonstrates visually the relationship between the visual and memory ratings tasks and identifications. Graphs depict ratings and identifications for all eight lineup members. Procedure was collapsed in these graphs.

Discussion (Studies 3 & 4)

The general aim of this study was to continue the exploration of similarity ratings initiated in study 1, and to determine whether similarity underlies identifications and can act as an alternative measure to identity for gauging eyewitness judgments. More specifically, the focus was on exploring (1) whether there were differences between the SIML and SEQL procedure in regards to resemblance ratings, (2) whether there were differences between perceptual and ecphoric resemblance ratings and identifications, and which was more accurate, and (3) the impact of the order of presentation, presence/absence of the target, and the ethnicity and gender of participants upon ratings.

Further evidence for similarity judgments underlying identifications

The results of studies 3 and 4 replicated those of studies 1 and 2 in providing evidence for a relationship between ratings and identifications. Persons rated clearly most similar were identified most often 100% of the time, and this was the case across visual and memory tasks and SIML and SEQL procedures. Whether a person was rated clearly most similar or not, in 91% of cases (or 29/32) there was a high or moderate level of match between ratings and identifications, meaning that the person rated most similar (even by a small degree) was the first or second most identified. This represented an increase from studies 1 and 2 where overall match was moderate or high in 72% of cases. The level of differentiation predicted the
level of match in 59% of cases, which was a decrease from studies 1 and 2 (69%). The purpose of this measure was to better incorporate situations where match would be expected to be low (e.g. no lineup member stood out). However, in a large number of cases, even when no lineup member stood out (Low-D), there was a moderate level of match, resulting in the reduced accurate prediction rate in studies 3 and 4 compared to studies 1 and 2.

**Are ratings more accurate than identifications?**

In short, SIML ratings were less accurate than identifications, while SEQL ratings were equivalently accurate to identifications. Accuracy (target identified or rated most similar) occurred in the following order: SEQL visual task, SIML visual task, SEQL identification task, SIML identification task, SEQL memory task, SIML memory task. Given that the visual task involved participants being able to compare two photos of the target it was expected that they would be most accurate. If anything, it was surprising that they were not accurate 100% of the time (SIML = 84.83%, SEQL = 96.75%). More important was comparing ratings and identifications in relation to the memory task. With regards to memory-based ratings, SEQL targets were rated most similar at an equivalent rate to them being identified, suggesting that overall, ratings were equally as accurate as identifications. When examined in more detail, SEQL identifications were more accurate than SEQL ratings in three out of the four target present lineups. Lineup 2 involved the target being rated most similar 73.9% of the time, and identified 44.4% of the time, and this may have inflated the overall accuracy of SEQL ratings. In comparison, SIML targets were rated most similar around 16% less often than they were identified. When examined in more detail this pattern was reasonably consistent across three target present lineups, though SIML ratings were 4.7% more accurate than SIML identifications in relation to lineup 5. This result involving the SIML identifications being more accurate than ratings replicated the trend observed in studies 1 and 2, though it was more pronounced in studies 3 and 4. There are several explanations for why the SEQL was more accurate than the SIML. These are explored in the section below comparing the two procedures. Discrepancy scores demonstrated a clear association between increased discrepancy and correct decisions in the memory tasks. This suggests that when greater ecphoric similarity was perceived the rating was more diagnostic of accuracy. It is also possible that this indicates that when less perceptual similarity was perceived between the first and second most highly rated lineup members (producing a larger discrepancy score) the highest rating was more diagnostic of accuracy.
SIML and SEQL procedure: Ratings

When the target was present the SEQL outperformed the SIML in regards to rating the target most similar more often, and this was the case across conditions and lineups. For instance, the SEQL resulted in the target being rated most similar 12-13% more often across both visual and memory tasks. In the memory task targets were rated most similar more often in the SEQL compared to SIML across all four target present lineups. Discrepancy scores were also higher in the SEQL compared to SIML across both visual and memory tasks, though the comparison was only significant for the former. This suggests that the reason targets were rated most similar more often in the SEQL was because participants perceived more similarity between the two photos of the target, and less similarity between the target photo and remaining lineup members in the SEQL compared to the SIML. That is to say that participants completing the SEQL appeared to make more nuanced similarity judgments. Thirdly, the SEQL ratings better approximated identification decisions. Thus the SEQL outperformed the SIML in three important areas.

Reasons for the SEQL outperforming the SIML with regards to ratings.

One interpretation for the SEQL outperforming the SIML is that the SEQL procedure facilitated better discrimination between faces. This is supported by the higher discrepancy scores in the SEQL, which suggest that not only was the target rated more similar, but other lineup members were rated less similar. This improved ability to discriminate may have been caused by the SEQL facilitating greater focus of attention for making comparisons between faces (be they perceptual or ecphoric comparisons). For instance, SEQL comparisons involved fewer faces and may have thus allowed a purer appraisal of ecphoric similarity (between one lineup member and the participant’s memory of the target). Along similar lines, the greater number of faces presented at once in the SIML may have been distracting, and subtracted from participant’s ability to make detailed judgments for each face. This explanation suggests that the SIML process itself can interfere with the process of making similarity judgments. That in the SIML visual task targets presented early were rated most similar more often (91.4%) than those presented late (78.3%) supports this interference explanation (this pattern was replicated in the memory task). Alternatively, the multiple faces in the SIML may have triggered the use of cognitive heuristics that allowed participants to manage to process the larger amount of information rapidly. If this was the case then the greater number of faces in the SIML may have encouraged some participants to scan the faces in a more
superficial manner. This would explain the reduced discrepancy scores in the SIML. However, it is important to acknowledge that while cognitive and attentional processes may have affected the completion of the eight lineups in this research, and lead to the SEQL procedure outperforming the SIML with regards to ratings, in practice they would be less likely to influence a real witness’s decision. This is because the gravity of the situation would likely over-ride factors pertaining to inattention, and inhibit heuristic processing.

Another interpretation is that the SEQL encouraged greater reliance on ecphoric processes (‘is this person similar to my memory for the target?’), whereas the SIML may have allowed for greater influence of perceptual processes (‘is this person visibly dissimilar to the other lineup members?’). Reliance on ecphoric similarity processes (rather than perceptual) would increase the accuracy of a rating, as this better addresses the question being asked of the witness: the witness compares their memory of the target to each lineup member, rather than comparing between lineup members. The SEQL encouraged ecphoric processes because presenting only one lineup photo at a time reduces the capacity for inter-lineup comparisons, thereby forcing greater reliance on making a comparison involving memory of the suspect. Thus, the SEQL may have better facilitated the question being asked of the witness, where the SIML allowed for more extraneous influence.

Another explanation for the SEQL outperforming the SIML with regards to ratings is that the combination of the SEQL with ratings formed a new hybrid lineup procedure that incorporated strengths of both the SIML and SEQL. For instance, the combination of the SEQL with ratings overcame a major problem of the SEQL identification procedure: that not all photos are necessarily seen. Using ratings with the SEQL meant that participants were aware of how many photos they would view, and were able to view them all. Other advantages of the SEQL, such as multiple questions, were maintained. Thus the SEQL with ratings may offer a powerful hybrid procedure that combines the strengths of the SIML and SEQL procedures, whilst overcoming some of their limitations.

That the SEQL outperformed the SIML regarding ratings of targets was surprising, as the SEQL involved a longer retention interval between viewing the target initially and subsequently at the lineup stage. Comparatively, in the SIML, all lineup faces were viewed shortly after initially seeing the target, and all at once. It was thought that this might impact upon the SEQL’s performance negatively and lead to a reduced hit rate compared to the SIML, which was not the case. The greater retention interval in the SEQL would not present an obstacle in practice, as
real life retention intervals are much greater, rendering this difference between the SEQL and SIML less meaningful. It was also considered possible that there might be greater memory interference in the SEQL from participants being forced to examine one face at a time, and that this greater effort would make it more prone to the effects of fatigue. However, surprisingly, while there was a definite decrease in participants’ performance over the course of the eight lineups, it was slightly more pronounced in the SIML condition.

**SIML and SEQL procedure: Identifications**

In regards to identifications, when the target was present, there was little difference in relation to lineup procedure: hits, identifications of foils, and misses were roughly equivalent across the SIML and SEQL. However, when the target was absent, the SIML resulted in slightly more correct rejections (57.43%) than the SEQL (49.08%). While not significant, this trend was in contrast with previous research, which has typically suggested that the SEQL is less diagnostic of guilt, but more diagnostic of innocence compared to the SIML. Diagnosis of innocence can be inferred from increased correct rejections or decreased false positives. Regarding the latter, most research has only included selection of the target replacement as false positive, with the rationale being that all other identifications can be immediately recognised as false in a one-suspect lineup. However, in the SEQL it is possible for a foil to be identified prior to the target replacement being observed in a target absent lineup. Presumably a target replacement is included in a lineup due to being most like the target. Given that not all faces are seen in the SEQL it may be that there is a greater likelihood of a target replacement being identified in the SIML rather than SEQL, leading to an inflation in the false positive measure for SIML lineups compared to SEQL. The SEQL would be least likely to result in a target replacement selection when target replacements are not highly distinctive, foils are highly similar, or the target replacement is presented late. This would then artificially inflate the capacity of the SEQL to diagnose innocence relative to the SIML. Thus, differences between studies can partly be attributed to the various definitions given to a false positive. For instance, if this study incorporated only selection of target replacements as a false positive, then it too would have found that the SEQL was more diagnostic of innocence in regards to false positives: target replacements were identified at a higher rate in SIML lineups (16.65%) compared to SEQL lineups (12.95%), but foils were identified more often in SEQL (37.95%) compared to SIML (25.9%) lineups. Thus, the difference in false positives in target absent lineups was driven by the identification of foils rather than target replacements.
However, this does not explain why in the current study the SIML lead to a higher rate of correct rejections. The SEQL has typically been found to result in a lower choosing rate. This has been explained according to either the relative/absolute distinction, or an internal criterion shift. The absolute/relative explanation posits that the nature of the SEQL forces a more definite decision making process, where each lineup member is compared to the participant’s memory of the target. Whereas in the SIML, the participant selects the lineup member who most resembles their memory of the target. That a person is not always selected in the SIML appears to present an obstacle to this account. For instance, in the current study, 16.65% of target present and 57.43% of target absent SIML lineups resulted in non-selection. According to the criterion shift explanation, the SEQL procedure is considered to result in a more conservative threshold for choosing, leading to decreased identification of targets in target present lineups, and increased rejection of target absent lineups. In the current study participants completing the SEQL made a selection in 80.6% of target present lineups, and 51% of target absent lineups. In comparison, participants completing the SIML made a selection in 83.4% of target present lineups, and 42.6% of target absent lineups. Why the SEQL procedure did not result in a more conservative threshold for choosing in the target absent condition is unclear. In all conditions, participants were advised on the informed consent page at the beginning of the study, that the target may or may not be present in the lineups. However, the SEQL took longer to complete, meaning that potentially participants completing it were more likely to forget this information. This may have rendered their threshold for choosing less conservative.

**Similarity is subjective**

One notable feature of similarity ratings in this study was their diversity between participants. Even in the visual condition where participant’s ratings were made between two present stimuli, perceptions of similarity represented by ratings differed notably. This was evident in that 15% of participants in the SIML visual condition did not rate the first photo of the target as most resembling the second photo of the target, despite both being presented simultaneously. One explanation for this is that these participants experienced inattention or low motivation in relation to the task. However, this seems unlikely as all participants completed the eight lineups when they were able to abandon the task at any time. More likely is that this reflects the difficulty of picking out a face from an array even when provided with a different photo of the same face. As demonstrated by Bruce et al. (1999) this
process is not automatic or universal, particularly when the pictures differ in expression, angle or clothing: in this study they differed in terms of clothing. There were also few ‘All 1 responses’ that might indicate a lack of attention or effort. More likely is that this finding provides a striking reminder of the inherent difficulties involved in providing eyewitness testimony and highlights the subjective nature of similarity. While no check task was included in this research to ensure that participants were not responding at random, this appears unlikely given the pattern of results – no participants in the final data set provided ‘All 1 responses’ across all lineups, and such responses were not more prevalent at the end than the start of the eight lineups. Furthermore, disinterested participants were more likely to drop out of the study than complete the eight lineups involved. The results of the visual condition indicate that even with a basic task involving perceptual similarity, people can interpret and perceive the similarity between two present objects quite differently. This finding emphasises that similarity is relative – what is regarded as similar differs between individuals. The subjective nature of similarity, and the implications for the subjective bias that a witness brings to the identification process, should not be discounted.

**Visual versus memory conditions**

Two questions of interest were explored in relation to visual and memory tasks. First, whether ratings of targets based on perceptual judgments were more accurate than those based on ecphoric judgments, and second, whether perceptual judgments were more closely associated with identification patterns than ecphoric judgments. It is important to keep in mind that perceptual judgments refer to judgments between two present stimuli. Thus they can refer to inter-lineup comparisons, which present an unwanted influence and are not related to accuracy. Alternatively, they can refer (as here) to a comparison between two photos of the target that were presented together. The latter should be highly correlated with accuracy. Regarding the first question, as expected, the visual task resulted in the target being rated most similar more often than in the memory task, and this was the case across both SIML and SEQL procedures. The average ratings of targets in the visual task were also higher, and there were significantly greater discrepancies between the ratings of targets and the next highest rating compared to the memory task. This suggests that perceptual similarity judgments involving two photos of the target were more nuanced and more accurate than ecphoric judgments. Furthermore, perceptual judgments facilitated a greater range of ratings for lineup members other than the target, compared to ecphoric ratings. Thus, as expected,
perceptual judgments made between two present stimuli allowed for a purer appraisal of the similarity/dissimilarity between them that was not hindered by factors affecting memory. This purer appraisal of similarity made possible using perceptual judgments resulted in more accurate ratings. This finding provides preliminary evidence that perceptual processes are a more basic building block of similarity that underlie ecphoric judgments.

Regarding the second question, the unexpected result was that there was little difference between the visual and memory tasks and identifications. The results of differentiation and match indicated that the relationship between ratings and identifications was similar across visual and memory tasks. In some ways this finding was unsurprising given that the memory ratings task more closely mirrored the identification task in general: both relied on ecphoric similarity processes comparing present stimuli and participants’ memory of the target. Thus one interpretation is that while perceptual factors do in fact underlie ecphoric judgments (evident in visual task resulting in higher ratings and larger discrepancies), this relationship was mediated by the semblance between the two memory based tasks. Both the latter relied on ecphoric similarity processes comparing present stimuli and participants’ memory of the target, and both were subject to the same factors affecting memory, which may have overridden any perceptual advantage. Thus, it may be that perceptual factors do indeed underlie ecphoric factors, but that a relationship will also exist between two ecphoric tasks, which are subject to the same interfering factors (e.g. memory decay). Had a visual identification task had been included (e.g. where the target was presented with the lineup as was the case in the visual ratings task), then the visual ratings and visual identification patterns would likely more closely mirror each other.

It also needs to noted that it was not possible to determine with certainty whether a participant was relying on exclusively perceptual factors in the perceptual task, or exclusively ecphoric factors in the ecphoric task. Namely the task was manipulated and it was assumed that a certain type of processing would ensue. It seems safer to assume that when the target face is presented with the lineup that a perceptual (visual) comparison will occur. However, it is more difficult to remove the influence of perceptual factors upon ecphoric decisions, and this would more likely occur with the multiple faces in the SIML. This might partly explain why the SEQL was more accurate with ratings across both visual and memory tasks: there was less opportunity for extraneous perceptual comparisons to occur.
Identity Crisis in Identification Evidence

Order effects: Ratings

That the SIML procedure would be more susceptible to order effects using ratings was not expected. More surprisingly, the order effect was only evident in the SIML visual condition where participants could observe all photos at the one time. Two explanations are relevant here, first, for the presence of a SIML order effect, and second, for the absence of SEQL order effect. Regarding the former, one explanation for targets being rated most similar more often when presented early in the SIML visual task is that participants were better able to determine a baseline for similarity: e.g. participants were able to provide a high rating to the early target, and lower ratings to subsequent persons. This assumes that participants progressed from left to right when completing ratings. Conflicting with this explanation, of the eight cases where targets were out-rated in the SIML visual condition, six entailed the highest rating being afforded to a lineup member to the right of, or following, the target. Further detracting from this explanation was that the finding was not replicated in the SIML memory condition, and an opposite trend was observed in study 1. As such, this finding may be more a reflection of idiosyncrasies in the SIML visual task participant sample. For instance, four participants were responsible for the eight cases where the target was out-rated. These participants may have been generally poorer at the task due to such factors as weaker eyesight, increased fatigue, more distractions in the environment, or reduced motivation. Thus, the small sample sizes may have played a role.

A second explanation is needed regarding the absence of order effects in the SEQL ratings conditions, particularly when an order effect was observed in SEQL identification condition. Order effects regarding SEQL identifications are likely due to the fact that not all faces are necessarily seen. This then puts greater emphasis on whether the target is presented early or late, as if the target is presented late, but a similar foil is presented early, then this will increase false positives of the similar foil. Inversely, if the target is presented early, then they are more likely to be identified, as was the case in this study. Foils were identified prior to the target being viewed in 8.3% of target present lineups, and prior to the target replacement being viewed in 24.1% of target absent lineups. This means that the target was sometimes not even seen in the SEQL lineup, which is concerning for the accuracy of the SEQL. Given that the ratings procedure involved all faces being viewed, it is not surprising that order effects were minimised. Furthermore, participants in the ratings condition were aware how many photos they would view and were not placed under pressure to make a selection as was the case in the
identification condition. Thus order effects in the SEQL may largely be a byproduct of the identification process.

**Order effects: Identifications**

In regards to identifications, order effects were more pronounced in the SEQL than the SIML, as was expected. In the SEQL, significantly more hits occurred when the target was presented early. There was a non-significant trend in the SEQL for foils to be identified more often when the target was presented late, and for more correct rejections to occur when the target replacement was presented early. There were more misses in target present lineups when the target was presented late. Clarke and Davey (2005) found that participants were more likely to make a selection early in a SEQL, which they attributed to participants ‘spending’ their identification on a similar looking foil prior to observing the target (e.g. a false positive). They suggested that the SEQL advantage was restricted to lineups where the target preceded the NB alternative. Consistent with this, the main order effect found in this study pertained to increased correct identifications of targets (rather than false positives) when presented early. When the target was presented late the number of false positives doubled in the SEQL, but remained constant in the SIML. Thus, it would seem that multiple order effects can exist and are affected by whether the target is present/absent, whether a similar looking foil appears prior to the target, and potentially other factors such as the overall level of similarity between lineup members and the target. For instance, if a target is presented late and a similar looking foil appears early, then this may result in increased early identification of foils. However, if the target is presented early, then this may result in more correct identifications compared to when the target is presented late (as was found in this study). These order effects have significant implications for practice, where the decision of where to position the suspect/target becomes highly important. Given that it cannot be known whether the suspect is innocent or guilty this is highly problematic and involves inherent bias: if an innocent suspect is presented early compared to late then there may be an increased likelihood of an incorrect identification; if a guilty suspect is presented early then there may be an increased likelihood of correct identification.

Regarding the current study, the reason fewer hits occurred in the SEQL when the target was presented late is likely due to three factors: first, late presentation involved viewing more faces prior to the target, which may have increased memory interference; second, there was an increased retention interval between viewing the target and making an identification, which may have increased
decay in memory of the target; and third, late presentation of the target provided more opportunity for a foil to be identified as five foils were presented prior to the target being observed, compared to two foils in the early condition. In target absent lineups, there was a trend in both the SEQL and SIML for more foils to be identified late in the lineup. This may have been the result of participants feeling increasing pressure to make an identification as the lineup approached its end. Thus it appears that both response criterion (pressure to choose at the end of the lineup leading to reduced criterion) and discrimination (greater interference from increased time/faces) factors are behind the presence of greater order effects in the SEQL. It may be that two internal thresholds are in operation, one pertaining to similarity, the other dissimilarity. These thresholds may change over the course of the lineup in response to levels of similarity/dissimilarity between lineup members.

As noted earlier, the ‘viper’ video system employed in England and Wales minimises the role of order effects by allowing witness’s two observational laps before making a selection, and they are able to return to faces in any order. However, there is still reason to believe that a person’s internal response criterion for choosing is impacted by the order in which faces are viewed. For instance, if the first face seen is most unlike the suspect this might lead to the witness adopting a less conservative criterion for choosing. This same phenomenon would also influence SEQL ratings, where it might be more likely to result in higher ratings of similarity being provided for lineup members who more resemble the target.

**Discrepancy as a measure of accuracy**

Discrepancy scores proved highly useful for exploring in more detail differences across the four ratings conditions (SIML-visual, SIML-memory, SEQL-visual, SEQL-memory tasks). Firstly, a clear relationship was evident with greater discrepancy between the highest and second highest rating being associated with increased accuracy. The SEQL procedure resulted in greater discrepancies than the SIML in both visual and memory tasks, though the difference was only significant in relation to the visual task. This may indicate that participants made more nuanced judgments in the SEQL compared to the SIML, both in terms of perceiving targets as more similar, and remaining lineup members as less similar. That the visual task resulted in significantly higher discrepancies than the memory task supports the notion that higher discrepancy scores index more informed decisions.

Low discrepancy scores were also useful for indexing target absent lineups, where the majority (76.4%) of the difference between the highest and second
The highest rating was 0 or 1. Thus, low discrepancy scores can be considered as one factor warning that the witness has not made a confident or informed decision. This is supported by the fact that discrepancy scores were much higher in all target present compared to target absent lineups. Discrepancy may therefore provide a useful measure of an eyewitness’s accuracy. A note of warning is appropriate, however, as one participant in the SEQL condition had the maximum discrepancy of 6 between the highest and second highest rating, but failed to rate the target most similar. Thus, discrepancy as a measure of accuracy is not free from error. The warning aside, information regarding discrepancy is relevant to the courtroom, and may provide a better index of accuracy than high ratings. Discrepancy might prove particularly useful if used in conjunction with other factors known to be associated with accuracy. For example, a witness’s evidence might be described in terms of their estimated distance from the suspect, whether the suspect was known to them, and the discrepancy between the highest and second highest rating, amongst other factors. This further allows jurors to appraise eyewitness evidence, rather than simply be provided with a binary outcome. Further more, discrepancy is an intuitive concept that can be easily understood by jurors and incorporated into their decision making process. It is worth noting, however, that the discrepancy/accuracy relationship will need to be considered carefully in order to prevent similar issues to identifications (e.g. a high discrepancy indicating ‘recognition’). How jurors might incorporate this information in their decision-making is explored further in the ‘Advantages of similarity ratings’ section, however, clearly some research is needed to examine mock jurors’ experience of integrating this information into their determinations.

A final benefit of discrepancy scores is that they can allow for probabilistic statements regarding the likelihood of a witness being accurate given the discrepancy between their highest and second highest rating. For instance, it was possible to state in relation to the current study that a discrepancy score of three or more resulted in the target being rated most similar 94% of the time (collapsed across procedure). In the same way that it is possible for certain persons to provide higher confidence ratings due to being more confident overall, it is possible some individuals might be more prone to providing high ratings than others. That is to say that there is a subjective component involved which needs to be accounted for. However, in practice, it might be possible to gauge a witness’s individual ratings by asking them to complete standardised practice lineups following their formal lineup rating/identification. This would provide information on their particular ratings pattern
and place them in the context of other raters. If, for instance, they provide high discrepancies for incorrect decisions, then less weight might be given to discrepancy scores for that witness. Exploring such a procedure was not within the scope of this research, thus the above comment is made with caution.

Brewer and his colleagues, who have performed similar discrepancy analyses in relation to confidence ratings (Sauer, Brewer & Weber 2012) appear to rely on an algorithm to determine a cut off point for accuracy. For instance, this would allow them to state that at a confidence rating of 92%, 85% of ratings were accurate. The algorithm presumably relies upon the input of data for both the lineup and the individual rater. But given that multiple data points appear necessary in order for an algorithm to function (e.g. multiple persons rating the lineup and the witness rating several lineups), which is not feasible in the real world, it is unclear how the algorithm applies to the lineup process in practice.

**Target present and target absent ratings**

Average ratings of targets were significantly higher than average ratings of the person rated most similar in the target absent condition. This was the case for all four ratings conditions. This demonstrates that ratings can to some degree accurately distinguish between targets and the highest-rated person in a target absent lineup. However, this information has limited value in diagnosing perpetrators from non-perpetrators, as on an individual participant level, persons in the target absent condition regularly received high ratings. Furthermore, in practice it is unknown whether a lineup includes the perpetrator or not. Thus, the fact that a person was rated highly is not diagnostic on an individual level of them being a target/perpetrator.

The current study examined the upper threshold of similarity, in relation to who was rated most similar in a lineup. However, there may be two processes involved, one for selecting and one for rejecting a lineup. An analogy for this is the two approaches to solving a multiple choice problem. This can be done by searching for the correct answer (selecting the most similar lineup member), or alternatively, by ruling out answers known to be incorrect using the cross out method (rejecting less similar lineup members). If the latter method is used, it would seem that there is more opportunity for a target replacement (innocent person) to be identified or rated most similar. This could be further explored by investigating the basis upon which people make their ratings. Ways to do this are discussed in the ‘Future Directions’ section.
The impact of gender and ethnicity

Some findings in relation to gender were mirrored in both ratings and identification conditions. Most striking was the interaction between gender and procedure. Males rated targets most similar more often in the SIML compared to SEQL, and identified targets more often in the SIML (though the sample was small and the result non-significant). Females rated targets most similar significantly more often in the SEQL compared to SIML, and also identified targets more often in the SEQL (although the latter was not significant). Thus, the significant interaction in regards to ratings was replicated in non-significant fashion with regard to identifications. That males performed slightly better in relation to targets in the SIML is consistent with previous research suggesting some bias towards identifying faces of one’s own gender (Wright & Sladden, 2003). However, why similar results did not occur in the SEQL is unclear. In fact the opposite was the case, with females being significantly better at rating targets most similar in the SEQL. The reasons for this finding are unclear. Even more strangely, males in the SEQL rated target replacements most similar significantly more often than females. Thus, the difference in gender in relation to the SEQL appears to be more related to better ability to discriminate rather than a preparedness to provide high ratings. The reasons for the interaction between gender and procedure require further exploration involving a larger sample. As noted, only four males completed the SIML condition in study 3, meaning that effects involving gender were interpreted with some caution.

No significant effects involving participant ethnicity were observed for ratings or identifications. However, the expectation for a bias towards own race was evident in the direction of rating results, which indicated that Caucasian participants rated targets most similar slightly more often than non-Caucasian participants. It is possible that the subtle affects of participants being able to make more nuanced judgments based on own-race bias were not captured by the 7-point Likert-scale employed in this study. In contrast to the finding that Caucasian participants rated targets most similar slightly more often, non-Caucasian participants produced slightly higher discrepancy scores for target present lineups compared to Caucasian participants. This might suggest that non-Caucasian participants made equally or finer discriminations as Caucasian participants in relation to Caucasian faces. However, whether larger discrepancy scores index more nuanced judgments, or even the opposite, cannot be determined using the current design. For instance, it is
possible that large discrepancy scores actually reflect the presence of blunted judgments made on the basis of reduced capacity to discriminate.

CHAPTER 11.

General Discussion

The aim of this research was to examine whether similarity processes underlie identifications in relation to unfamiliar faces, and to explore the potential for similarity ratings as an alternative to identifications in eyewitness evidence. This involved exploring two questions, first, whether a relationship existed, and second, whether ratings were diagnostic of accuracy (and therefore useful). Given the exploratory nature of this research only a descriptive comparison of similarity ratings and identification data was presented. With this limitation in mind, results provided strong preliminary evidence for the presence of a relationship between similarity ratings and identifications. Results also indicated that ratings were equivalently accurate to identifications in relation to the SEQL procedure, but that identifications were more accurate than SIML ratings. The theoretical and practical advantages of ratings are discussed as well as their limitations. Recommendations for future research and applications are provided.

The existence of a relationship between similarity ratings and identifications: Preliminary evidence obtained

The level of relationship was gauged using the concepts of differentiation and match that were developed in this research. Differentiation referred to whether any one person in a lineup stood out with regards to ratings, and provided an indication of situations where a strong relationship was expected. High differentiation was expected to result in a stronger relationship between ratings and identifications. Comparatively, when there was a smaller difference between the highest and second highest rating (low differentiation), it was considered less likely that the highest rated person would be identified most often, and a diminished relationship was expected. Match referred to the level of agreement between the two types of data, with high match indicating that the person rated most similar was most commonly identified – evidence of a strong level of agreement. Moderate match referred to situations where the person rated most similar was identified

---

9 Differentiation was operationalised as follows: Low differentiation (Low-D) = less than or equal to 10% difference between the first and second persons rated most similar; Moderate differentiation (Mod-D) = between 10.1 - 20% difference; and High differentiation (High-D) = when the difference between the two exceeded 20%.
second most often, or where the person identified most often was rated second most similar.

The presence of a strong relationship between similarity ratings and identifications was evident in the fact that across both studies, a person who was rated as clearly most similar (highly differentiated) was identified most often in 100% of cases. In contrast, a person who did not stand out (low differentiation) in regards to ratings was identified most often in 20% of cases. Furthermore, there was a high or moderate level of agreement between ratings and identifications in 72% of cases in studies 1 and 2, and 91% of cases in studies 3 and 4. Thus, evidence for the existence of a relationship was obtained, and this relationship was stronger in certain circumstances.

That the relationship was strongest when one person was highly differentiated was hardly surprising, as for a relationship between similarity and identity to exist, there must be some measure of similarity present. Thus, the relationship between similarity ratings and identifications was dependent on some level of similarity being perceived as present. While ratings in the visual condition reflected perceptual similarity, ratings in the memory condition were assumed to reflect more ecphoric factors, though the possibility of some degree of perceptual influence could not be excluded. For instance, a person might be rated highly due to being highly similar to a participant’s memory of the target (ecphoric similarity), or due to being highly dissimilar to the remaining lineup members (perceptual dissimilarity). It is preferred that judgments rely on the former, as these would be more accurate. Ecphoric judgments also better address the question being asked of a witness – e.g. witnesses are not being asked whether a person is more similar than others in a lineup, but whether they are similar to the perpetrator.

Whilst evidence for the presence of a relationship between similarity ratings and identifications was observed, the direction of this relationship could not be confirmed using the current experimental design. For instance, whether similarity processes underpinned identification judgments or the reverse could not be ascertained. Additionally, the possibility cannot be ruled out that some third variable exists that explains this relationship. Nonetheless, logically, it appears more likely that basic similarity processes provide the basis upon which more complex judgments regarding identity are made, rather than the reverse. Support for this interpretation was obtained in study 3, which involved visual comparisons between two different photos of the target. Perceptual judgments involving target comparisons were more accurate than ecphoric (memory-based) judgments. One
interpretation of this is that perceptual judgments represent more basic building blocks of similarity upon which ephoric determinations are founded. However, further research is needed to confirm these initial results.

**Are similarity ratings diagnostic of accuracy?**

An important question in regards to ratings is ‘are they accurate?’ If not, then they are of little use. Accuracy was defined according whether the target was rated most similar or not. This measure of accuracy was preferred over average ratings as it was easily transformed into a binary outcome (e.g. target rated most similar Yes/No) rendering it comparable to identification data. Furthermore, average target ratings did not provide any reference to ratings for other lineup members, and did not account for the subjective nature of ratings (e.g. a person might rate the target most similar at 3/7). Finally, average ratings appeared to obscure the relationship between similarity and identifications, which the highest rating measure better illustrated.

In studies 1 and 2, which involved the SIML procedure, ratings were slightly less diagnostic of accuracy than identifications, with 40% of targets rated most similar, and 44% of targets identified. In studies 3 and 4, the equivalent memory-based SIML ratings were again less diagnostic of accuracy than identifications, with 53.6% of targets being rated most similar, and 70.4% being identified. However, the memory-based SEQL ratings were equivalently accurate to identifications, with 66.3% of targets being rated most similar, and 68.5% identified. Thus, the SEQL ratings performed on par with identifications, but SIML ratings were less accurate. In contrast, Sauer et al. (2012) found that making time-pressured confidence ratings regarding whether a face had been seen before were more accurate than identification decisions. However, there were a number of differences between that study and the present research, such as the time-pressured decision-making, use of the SEQL procedure only, and confidence ratings. A criticism of ratings based on the above findings might be that they do not offer an increase in accuracy over identifications, and are therefore of limited use. However, several reasons are explored why this is not the case. Reasons are discussed for (1) why SIML ratings were less accurate than identifications, and (2) why SEQL ratings were more accurate than SIML ratings.

**Reasons why SIML ratings were less accurate than identifications**

Several potential reasons why SIML ratings were less accurate than identifications exist. These explanations draw on the response criterion, differences
in facial processing between ratings and identification tasks, memory interference, and the fact that identifications can occur in the absence of similarity.

Response criterion explanation – threshold responding was better captured by ratings.

The first explanation for why identifications were more accurate than ratings involves the response criterion, which posits that for an identification to occur an internal decision threshold must be crossed. It is likely that some correct identifications were made with the criterion operating at threshold (on the cusp of making an identification). This means that some of the correct identifications were in effect educated guesses – participants made a selection when they were uncertain. The 12-15% of false positives in target present lineups may represent the flipside of these threshold decisions that were incorrect. Support for the hypothesis that some identifications were made at threshold comes from two places. First, foils were identified instead of the target in approximately 14% of cases, suggesting some level of over-preparedness to make a selection without necessary discriminatory capacity. Second, and more important, on average approximately 12% of ratings involved another lineup member being rated equally as highly as the target. An example of this includes the target and one or more lineup members being rated as 85% similar to the participant’s memory of the target. This means that in 12% of cases (averaged across both studies and procedures) multiple persons were perceived as equally similar to one’s memory for the target. A comparable number of ‘equal highest ratings’ occurred in the target absent condition - on average approximately 12% across both studies and procedures.

Thus, the argument is that whilst more SIML targets were identified than rated most similar, in both conditions there were instances where some level of uncertainty existed. Where with identifications some of this uncertainty was translated into a correct rejection of the lineup, sometimes it resulted in identifications, with some being hits and others false positives: the identification procedure does not allow for determining which. In comparison, with ratings, this uncertainty was captured as an ‘equal highest rating’ response. Put another way, such a pattern of ratings indexes a person’s uncertainty and may sound a warning regarding the reliability of associated eyewitness evidence. No such warning is provided with regard to identifications. Notably, making an identification with a response criterion operating at threshold can lead to an incorrect identification, however, it can also lead to incorrect rejections. In practice, a witness may fail to identify a suspect despite them being perceived as most similar, leading police to
exclude the suspect from further investigation. This could result in guilty persons evading detection. With ratings, if a suspect was regarded as being most similar, but not enough to trigger an identification, this would be captured, thereby increasing the likelihood of continued police investigation of the suspect.\textsuperscript{10} In this context similarity ratings offer greater transparency and provide valuable additional information and a more precise index of a persons’ memory of a target relative to other lineup members.

\textit{Identifications may have relied more on global/automatic/relational processing whereas ratings encouraged local/conscious/attributional processing.}

Another explanation for why correct identifications were higher is that there was some aspect inherent to the identification process that facilitated accurate decisions over ratings. For instance, ratings in this study may have triggered more laborious conscious cognitive processing where identifications relied more on automatic global judgments. The advantages of automatic global processing over more intentional local processing have already been described. A previous study that primed participants to employ global or local processes found that in a subsequent SIML task, those primed to use global judgments performed better than controls and or those primed to make local judgments (Macrae & Lewis, 2002). The size of the effect was striking, with facial recognition accuracy being 83\% for the global processing group, 60\% for a control group, and 30\% for the local processing group. Perfect (2003) replicated their study with similar results. Further supporting this interpretation, Sauer et al.’s (2012) study involving the SEQL required participants to provide confidence ratings within 3 seconds, which likely forced more global automatic processing. In contrast, the process of requiring participants to rate every lineup member in this study without time pressure may have encouraged more conscious feature-based processing in relation to ratings. In the identification condition participants were able to zone in on the most similar or familiar face thereby removing the distraction of the remaining faces. This would facilitate more automatic processing. This interpretation is supported by the fact that the effect was largely in relation to the SIML, where the larger number of faces presented at once would amplify the interference.

If global/local processing differences do in fact partly underlie the higher identification accuracy, then this suggests that the nature of the task can interfere

\textsuperscript{10} It is acknowledged that this was not the approach taken in the current research, where for the purpose of analyses ratings were converted into a binary outcome (rated most highly - Yes/No) for the sake of comparison with identification data.
with the remembering processes. Interestingly, Sauer et al. (2012) appeared to regard the increased accuracy for confidence ratings over identifications in their study as being more due to the deadline procedure (responding within 3 seconds) than the ratings process itself. They noted that if the witness had more time for reflection, then “previous exposure to one of the lineup foils might produce a sense of familiarity, that, on reflection, could be attributed to the face being that of the culprit.” (p. 1213). They then posed the question of whether the deadline procedure was a necessary component for increasing accuracy. They do not appear to have answered this question with further research as of yet. Some potential strategies for reducing the impact of global/local processing upon ratings are explored below.

In conflict with this explanation, not all research has found that priming global processing increases accuracy. Lawson (2007) replicated Macrae and Lewis’ (2002) study across three experiments, none of which showed any benefit for facial recognition following priming for global processing. Brand (2004) also failed to find effects anywhere near those reported by Macrae and Lewis. Lawson concluded that her results, taken together with those of Brand, suggested that the processing bias effect was at best weak and difficult to replicate. These differences in research findings may refer to whether a processing effect exists (e.g. whether global is superior), or the ability of the studies to successfully prime particular processes. For instance, it may be that Lawson and Brand’s research was less successful at priming global processes, leading to the absence of an effect. However, the studies reported relied on similar methods for priming particular processes (e.g. Navon stimuli) and included similar tasks, meaning that an inability to prime explanation appears less likely.

A separate explanation that also emphasises a difference in processing as the reason for differences in accuracy is provided by past research. One previous study examined similarity and dissimilarity judgments for objects separated between two types of similarity judgments: attributional and relational (Medin et al., 1990). Attributions refer to any constituent property of a stimulus, where as relations refer to descriptions of connections between two or more objects or attributes. The authors gave the example where stimulus ‘A’ consists of a red square and a red circle and stimulus ‘B’ consists of a blue circle and a blue triangle. The A and B stimuli would share the attribute ‘circle’ and the relation ‘same colour’. The authors concluded on the basis of their results that either (a) attributional matches were less important than relational matches for judgments of similarity, and/or that (b) attributional mis-matches were more important than relational mismatches in
dissimilarity judgments (their research design did not discriminate between these two possibilities). They noted that although sameness judgments are typically described as more global or nonanalytic than difference judgments, an alternative possibility is that they focus on relations rather than attributes. Relating this to the current research, it might be that identifications relied more on relational similarity judgments, and ratings relied more on attributional similarity judgments, explaining why the former were often more accurate. Ratings may have triggered attributional processing due to the requirement to rate every face – a complex task that would be simplified by focusing on one or more attributes. Further, ratings may have primed attributional processing due to the requirement to consciously make dissimilarity judgments for the less similar faces (e.g. persons 1 and 2 in the lineup were always foils). In comparison, identifications in the SIML allowed for faster and more efficient ‘zoning’ in to the person most representing the target, and thus better facilitated fewer relational similarity judgments. Potential ways to manage this issue regarding ratings are explored later in this chapter.

*Rating every lineup member may have increased memory interference.*

Rating every lineup member may have increased the level of memory interference experienced by participants. This could be in relation to the requirement to focus on each face, or the increased retention interval between viewing the target in the observation and lineup stage in the SIML. Regarding the former, the assumption is often made that the faces previously viewed in a lineup do not contaminate the process of viewing the current face and comparing it to the target. More precisely, it is assumed that faces 1-4 in the lineup do not affect the comparison between face 5 and a witness’s memory of the target. Whether this assumption is valid is unclear. However, one explanation for the lower rate of targets being rated most similar compared to identified in the SIML is that the ratings process increased memory interference due to the requirement for participants to focus on every face. The ‘verbal overshadowing effect’ (Schooler & Engstler-Schooler, 1990) refers to the negative effect whereby verbal descriptions impair visual recognition: witnesses who provide verbal descriptions are less able to recognize a suspect’s identity in a lineup. In one of a series of experiments, Schooler and Engstler-Schooler found that limiting participants’ time to make recognition decisions alleviated the verbal overshadowing effect, suggesting that it overshadows, rather than eradicates, the original visual memory. They concluded that verbalising a visual memory potentially produces a verbally biased memory representation than can interfere with the application of the original visual memory.
It is possible that requiring participants to rate every person in the lineup actually impaired their memory for the target and created a visual overshadowing effect. Making a rating involves comparing a lineup face to one’s memory of the target. Thus it is possible that repeatedly doing so has the consequence of impairing the memory of the target to some degree. Such an effect might be more likely to occur when the target was unfamiliar and seen only for a short time, as was the case in this research. This is because the representation of the target in participant’s memory was less strongly imprinted and thus less resilient to contamination. Such an effect might also be stronger in relation to the SIML because more perceptual comparisons are possible: e.g. person 1 can be compared to persons 2 to 8, as well as the participant’s memory of the target; in comparison, it is more difficult in the SEQL for person 1 to be perceptually compared to persons 2 to 8.

The need to rate every face would also have resulted in a longer retention interval between viewing the target initially and in the lineup. This would have increased decay of participants’ memory for the target. This effect would be expected to be stronger for the SIML (as was the case), as the SEQL procedure already involves a longer retention interval due the requirement to view one face at a time. Consistent with this, while targets presented late in the SIML were identified more often than those presented early (early = 66.7%, late = 74.1%), the reverse was true with regards to ratings. Across both visual and memory SIML ratings tasks targets presented early were rated most similar more often than those presented late (visual task early = 91.3%, visual task late = 78.3%; memory task early = 57.15%, memory task late = 50%). This suggests that memory decay may have played a role in why SIML ratings were less accurate than SIML identifications. Importantly, the impact of retention interval and memory decay are likely to be artifices of the experimental design. In actual practice, the naturally occurring longer retention interval and infinite number of faces seen would minimise this difference between procedures and conditions. The issue of memory interference and some potential strategies for managing it are discussed below.

*Identifications do not necessarily depend on similarity.*

It was suggested in the introduction that identity does not necessarily depend on similarity. For instance, a person may be identified on the basis of feelings of familiarity rather than visual similarity. Or change in certain attributes may reduce the perception of similarity whilst still allowing for the person to be identified. This would particularly be the case if identifications did indeed draw on more automatic/holistic judgments and ratings upon more conscious/localised judgments.
For example, significant changes in background as well as subtle changes in hairstyle, expression and lighting were apparent between the two photos of the targets employed in this research. Thus, it is possible that some persons may have been identified who were not perceived as being highly visually similar to the target. That is to say some identifications potentially relied more on a sense of familiarity rather than visual similarity. It was expected that the use of unfamiliar faces would minimise reliance on ‘feelings of familiarity’, but some such processing may have remained even for faces seen only once previously for five seconds. Why this issue would impact SIML ratings more than SEQL ratings is not clear.

Perhaps at least two types of ‘familiarity’ can affect identifications: conscious and unconscious. The former could be easily identified by asking the witness whether the suspect is known to them. The latter may be more subtle and difficult to detect. This might involve an identification occurring due to a lineup member resembling another person, and thus fostering a less conscious form of familiarity. Regarding the latter, it might be possible to attempt to manage it by including some measure of familiarity. For example, the witness might be required to provide some measure of familiarity over and above that of similarity. Instructions to the witness might be as follows: “This process will involve three steps, first you will be asked to select those members from the lineup that are most likely to be the perpetrator based on your memory. You will then be asked to rate each of these in terms of how similar they are to your memory of the perpetrator. It is possible for a person to be the perpetrator while not appearing visually similar to them. Thus, you will also be asked to indicate a sense of how familiar each of the selected persons are to your memory of the perpetrator.” However, such a process is accompanied by risk. Familiarity is a two edged sword in that persons may be incorrectly identified/rated most similar on the basis of mistaken feelings of familiarity. As reported earlier, judgments of familiarity are incorrect almost 25% of the time (Pezdek & Stolzenberg, 2014). Thus, including a measure of familiarity might introduce a range of new errors as well as potentially confuse witnesses.

Requiring the witness to offer a statement regarding the basis of their ratings could provide some insight into the process (though the witness may struggle to articulate it), and potentially flag issues such as that above. For instance, the first step outlined immediately above does not explicitly refer to similarity, but those lineup members who are “the most likely” to be the perpetrator. The following selection would then include persons selected on the basis of feelings of familiarity, though they might receive a lower similarity rating. The third step outlined above
could be replaced with a question regarding the witness’s highest similarity rating: “Is this person the most likely to be the perpetrator?” A “no” response would highlight the above issue, and allow for further probing. Such a question still avoids asking the witness to make a statement regarding the suspect’s identity – though it does come very close. A final approach to this issue would be to accept it as a necessary evil of ratings: that in some instances an identification may occur without depending on similarity. Such instances are likely to be rare.

Ways to minimise the problems outlined above.

Even if ratings do indeed change the process by which eyewitnesses make decisions, and/or increase memory interference, with a side effect being a reduction in accuracy, ratings still offer a range of benefits over identifications. However, there is also a number of potential ways to minimise the problems outlined above. First, instructing witnesses to rate faces in order of similarity, from the most to least similar, would reduce the level of memory interference in relation to the target rating. This change would only benefit the SIML, where it would allow the witness to zone in on the most similar face immediately. Appropriately, the SIML is more in need of improved accuracy than the SEQL. This change is simple and easily made. Furthermore, it would help test the hypothesis of memory interference partly contributing to reducing ratings accuracy.

Second, it would be possible to encourage more global processing and minimise the level of conscious effort by requiring witnesses to only rate those lineup members who are in some way similar to their memory of the perpetrator. This would allow them to skip past less relevant (less similar) faces that are a distraction at a faster rate. This could also reduce memory interference. Faces that did not receive a rating could be automatically allocated a ‘1’ (‘very little resemblance’) response without this needing to be manually completed by the witness. It is likely that this change would benefit the SEQL more than the SIML. This is because witnesses completing the SIML would already have been encouraged to focus on the most similar faces first (courtesy of the first recommended change), thereby reducing the influence of this latter alteration. In contrast, the first recommended change is not applicable to the SEQL, meaning that allowing witnesses to waste less time and attention rating earlier lineup members who are of little resemblance to the perpetrator might be of significant benefit. This would involve skipping past less relevant faces in the SEQL without providing a rating.
A way of further reducing distraction in the SIML would be to create lineup functionality that allows persons who exceed a dissimilarity criterion (e.g. ‘Not at all similar’) to be removed from the lineup, leaving the remaining persons to be rated. For example, if a SIML lineup member receives a rating of ‘1’ (“Not at all similar”), then their photo would be blacked out. The same information is retained (e.g. that removed persons were equated with the label ‘1’), and the remaining ratings are less likely to be influenced by perceptual comparisons to blacked out lineup members. For instance, a witness might be less likely to think that person 1 is highly dissimilar to their memory of the perpetrator, but in comparison to person 1, person 2 is much less dissimilar, and so provide a higher rating based on dissimilarity rather than similarity. In other words, using the black out procedure, the remaining similarity ratings might be more ecphoric in relation to the target and less perceptual in relation to other lineup members. Whether such a procedure reduced ratings in target absent lineups would also be of interest.

An alternative, and slightly different approach to reducing the number of persons being rated might involve participants first being asked to select those members of the lineup that are most likely to be the perpetrator based on their memory. They would then only provide resemblance ratings for these persons, with others being allocated a ‘1’ response. The witness could then be asked of the highest rating (if there is one): “Is this person likely to be the perpetrator?” or “Is this person most likely to be the perpetrator?” The witness’s response could be captured on a three-point Likert-scale ranging from ‘1’ – “Unsure”, ‘2’ – “Possibly”, ‘3’ – “Quite Likely”. Note that this scale deliberately avoids a highly conclusive response (e.g. such as “Definitely”). Finally, the witness could complete a post-rating statement (either after or instead of the Likert-scale) that describes their judgment process: e.g. “Please provide a statement describing your decision process”. The statement would then be presented to the jury to provide further detail.

Allowing witnesses to provide their rating verbally rather than manually entering it into a computer would further reduce the amount of cognitive effort required to for them to record rating information. The encouragement to rate faces as quickly as possible might also help prime more global processes and encourage participants to focus mainly on those faces most similar to their memory of the target. Two recommendations might be combined, and participants asked to “Rate only the faces that are similar to your memory of the target. Do this as quickly as you can.” The suggestion involving speed is more to help determine whether global or local processes are at play, as it seems inappropriate that lineup instructions in
real life require a witness to rush their decision. Whether due to memory 
interference, global/local or attributional/relational processing, these amendments to 
the ratings process would help simplify the task of the witness, and potentially 
facilitate faster and more accurate ratings. Naturally, this will only be confirmed 
through further research that has implemented these recommendations.

**Reasons for why SIML ratings were less accurate than SEQL ratings**

Three reasons were provided for why the SEQL outperformed the SIML with 
regards to rating targets most similar in the study 3 and 4 combined discussion. The 
first was that, given that the SEQL limits comparisons to two faces at a time, it may 
allow for a greater focus of attention, thereby facilitating capacity to discriminate. In 
contrast, the greater number of faces presented at once in the SIML may have been 
distracting and impeded participant’s ability to make an informed judgment for each 
face. This interference explanation was supported by the fact that in study 3 both 
the SIML visual and memory tasks targets presented early were rated most similar 
more often than those presented late. The second was that the multiple faces in the 
SIML may have triggered the use of cognitive heuristics that allowed participants to 
manage to process the larger amount of information rapidly. The likelihood of this 
occurring might have increased due to participants being informed that they would 
complete eight lineups. According to this explanation the greater number of faces in 
the SIML encouraged some participants to scan the faces in a more superficial 
manner. This would explain the reduced discrepancy between the highest and 
second highest ratings in the SIML compared to SEQL. The third was that the SEQL 
encouraged greater reliance on ecphoric processes (‘is this person similar to my 
memory for the target?’), whereas the SIML allowed perceptual comparisons greater 
influence (‘is this person similar/dissimilar to other lineup members?’). The SIML 
procedure better facilitates perceptual-based inter-lineup comparisons due to the 
fact that all the faces can be viewed at once. In contrast, the SEQL makes such 
inter-lineup comparisons more difficult given that such comparisons must occur in 
the witness’s memory. Thus, the SEQL procedure may better encourage ecphoric-
based judgments compared to the SIML, and the SIML may better encourage 
perceptual-based judgments compared to the SEQL. Given that the question being 
asked of a witness pertains to ecphoric similarity, reliance on these processes would 
result in greater accuracy, which partly explains the SEQL procedure’s advantage in 
relation to ratings. Thus, the SEQL may have better facilitated the question being 
asked of the witness, where the SIML allowed for more extraneous influence. A 
further reason the SEQL outperformed the SIML in ratings is that the combination of
the SEQL with ratings represents a powerful new hybrid procedure for capturing eyewitness evidence. For instance, SEQL ratings incorporated the strengths of both the SIML and SEQL identification procedures whilst removing their limitations. This is explored further on page 127 below.

Another explanation for why the SEQL outperformed the SIML draws on the argument provided above regarding identifications facilitating more global automatic processes compared to ratings. This explanation can equally be applied to why the SEQL outperformed the SIML. Global processing would appear to be better facilitated by the SEQL, where one face is viewed at a time. This shares some similarities with, but is distinct from, a relative/absolute explanation. The relative/absolute explanation suggests that the problem with the SIML is that inter-lineup comparison results in the person who most approximates the target being selected, thereby increasing false positives. This is thought to be avoided using the SEQL procedure, which encourages a more definite and certain decision-making process. In contrast, the global automatic explanation suggests that the difficulty with the SIML is that it primes more laborious and conscious processes that focus on the parts rather than the whole, which can be less accurate than more automatic decisions based on global processing. There is evidence to suggest that participants make more accurate identification decisions when primed to make global rather than local judgments, though the findings are mixed. Global judgments rely on encoding the relationship between features rather than focusing on individual features. More global processes can be promoted by making participants respond within a certain timeframe (Brewer, Weber, Wootton & Lindsay, 2012), or by priming local or global processing in pre-identification activities (Macrae & Lewis, 2002; Perfect, Dennis, & Snell, 2007).

Advantages of similarity ratings

Ratings provide a more nuanced picture of a witness’s memory.

A number of advantages for ratings were realised in this research. First, ratings allowed for a more nuanced picture of how much similarity a witness perceived between each lineup member and their memory of the target than was afforded using identifications. This meant that how similar one person was perceived to be could be interpreted relative to the similarity of other lineups members – a considerable benefit that is not afforded by identifications. As noted ratings provided a better index of a witness’s certainty or uncertainty, evidenced by a non-suspect lineup member being rated most similar, the suspect being rated
equally similar as another lineup member, or the suspect being rated clearly more similar than anyone else. This more nuanced picture is both important in regards to the investigative stage, where police capture more information regarding a witness’s memory. But it is also important in regards to the courtroom, where the jury can better assess the evidence, the end result of which is more informed decision making.

*Ratings provide three useful questions for juries.*

How ratings might be employed in practice is as follows. Rather than being provided with a single suspect selection, the jury could be presented with some or all of: the lineup photos, the witness’s similarity ratings, the witness’s rating of the likelihood of the suspect being the perpetrator (e.g. ranging from ‘Unsure’ to ‘Quite Likely’), and the witness’s description of the basis for their ratings. Three questions might then be asked of the evidence: (1) was the suspect rated most similar? (2) was the suspect rated highly overall? (3) what was the discrepancy between the highest and second highest rating? The first question is not equivalent to an identification, and should not be considered as such. This is an important point, as the purpose of ratings is not to provide an identification by another method. However, for the suspect to not be rated most similar is powerful evidence that they are unlikely to be the perpetrator. Thus, for the suspect to be rated most similar provides a requisite condition that should be satisfied for the evidence to support a conviction. The second question provides some measure of the overall similarity of the suspect (presuming they are rated most similar). For instance, if the scale is 1-7 a highest rating of 2/7 (the rest being 1/7) is less persuasive than a highest rating of 6/7. In this research targets on average received significantly higher ratings than similar looking target replacements, suggesting that high ratings are more indicative of accuracy than low ratings. Clearly the subjective nature of similarity presents a problem here, as one witness may naturally have a lower baseline rating than another. However, information regarding a witness’s baseline could be obtained by requiring them to complete several standardised practice lineups following the real lineup. These standardised lineups would need to be normed on large samples, thereby enabling them to provide a reliable indicator of an individual’s baseline relative to others. Regarding the third question, results from this research showed a clear relationship between discrepancy scores and accuracy, with greater discrepancy between the highest and second highest rating being associated with increased accuracy. Discrepancy scores are useful both to give an indication of the perceived difference in similarity (e.g. a larger discrepancy is more persuasive), but
they also allow for probabilistic statements regarding the likelihood of a witness being accurate given the discrepancy between their highest and second highest rating. For instance, it was possible to state in relation to study 3 that a discrepancy score of three or more resulted in the target being rated most similar 94% of the time. Conversely, low discrepancy scores were indicative of the target being absent, with the majority of discrepancy scores (76.4%) being 0 or 1 in relation to target absent lineups. These probabilistic statements may be a powerful addition to current eyewitness evidence. There was one instance in this research where a foil was rated most similar with the maximum discrepancy, which serves as a warning that discrepancy scores are not free from error. However, as suggested, with standardised lineups it would become possible to indicate the likelihood of this error occurring. While these three questions are useful for assisting juries assess eyewitness evidence, they are also useful for assisting police in the investigative stage. Police might rely on these three questions to help them determine whether a suspect is worthy of further investigation or not.

Ratings better represent reality.

If a relationship between similarity and identity exists, and this research provides preliminary evidence for this, then it appears more likely that similarity processes underlie identifications rather than the reverse (though this could not be ascertained using the current design). If this is the case, then attempting to force a binary decision upon what occurs on a continuum distorts reality: it is artificial to suggest that a witness is capable of ‘identifying’ a perpetrator. In this context, ratings better accord with a witness’s experience. This means that both the collection and presentation of eyewitness evidence using ratings better represents reality, and thus contributes to a fairer and more just process.

The reduced import given to ratings may facilitate the collection of supporting evidence.

The use of ratings undoubtedly dilutes the power currently afforded identification evidence. For a witness to state that “this is the same person that I saw commit the offence” is very persuasive, and this is evident in the fact that the positive identification of a suspect by a single witness is considered sufficient to result in conviction. However, given the problems associated with using ‘identity’ as the measure by which eyewitness evidence is quantified, a reduction in the power of this form of evidence is appropriate. It may even result in greater emphasis being placed on the collection and presentation of supporting evidence.
besides the three questions juries should ask described above, they might be encouraged to turn their mind to a list of other relevant factors. These include the distance of observation, duration of observation, level of familiarity of the suspect, how many times seen previously, and in what context, and whether the suspect has an alibi. This approach is analogous to that incorporated by structured professional judgment tools, which are typically employed in a forensic context to assess some aspect of risk (e.g. violence, Webster, Douglas, Eaves, & Hart, 1997). This typically involves a list of factors pertinent to the type of risk being assessed, which do not sum in an additive fashion, but merely serve to ensure the assessor has considered all the relevant factors. An example of a list of questions that might be completed by witnesses and presented to jurors, as well as separate questions jurors should consider, are presented in Appendix D. A checklist like this would assist in ensuring relevant information is both collected and taken into account in judicial decision-making (the list provided does not include the confidence of the witness, or whether the witness’s description matched the accused as the relationship of these factors with accuracy is questionable – see Thomson, 2003, for a review). The consequence of using such standardised procedure would be to create a minimum standard of information required, and encourage greater reliance on other supporting evidence.

*SEQL ratings provide a powerful new hybrid procedure for capturing eyewitness evidence that overcomes major problems associated with SEQL identifications.*

One possible conclusion based on this research is that the marriage of the SEQL procedure with similarity ratings may represent a powerful new hybrid procedure for capturing eyewitness evidence. For instance, the SEQL outperforming the SIML with regards to ratings may reflect the fact that SEQL ratings incorporated the strengths of both the SIML and SEQL identification procedures whilst removing their limitations. Two major criticisms of the SEQL identification procedure are that not all photos are necessarily seen, and that it is susceptible to order effects (these criticisms apply to the Australian and American contexts but not the UK where the ‘viper’ system overcomes both). This also means that backloading (participants being unaware of the number of photos being presented) is necessary. In contrast, using ratings, participants were aware of how many photos they would view, and were able to observe all photos. Significantly, there were no order effects present with regards to SEQL ratings, despite such effects being present in the identification condition. This is an important result, as the presence of order effects cast a shadow...
over the legitimacy of the SEQL procedure: if positioning a suspect in a particular place within the lineup increases their chance of being selected, then to put them in such a position unfairly increases their likelihood of being identified, and to not put them in such a position may impact upon correct convictions (depending on whether the suspect is the perpetrator). This issue is particularly salient given the great weight afforded to eyewitness testimony by jurors. As such, the hybrid procedure of the SEQL with ratings may present an important advancement in eyewitness testimony processes. Further research is needed to confirm those results reported here.

That not all photos are necessarily seen using the SEQL identification procedure has practical and research-based implications. The practical implication is that a foil presented prior to a target/target replacement may be identified without the target or target replacement even being viewed. This occurred in 8.3% of target present and 24.1% of target absent lineups in study 2, which represents a real problem for the SEQL identification procedure in practice. Alternatively, a target may be identified, when a subsequent foil may have been identified if given the opportunity to be observed. The fact that not all faces in the SEQL are necessarily seen may partly explain why it typically performs worse than the SIML when the target is present, but better when the target is absent. When all faces are seen (as when a target absent lineup is rejected) the SEQL strengths become more apparent.

A research-based implication is that measures of accuracy often exclude incorrect identification of foils as false positives on the basis that they are a known error (Lindsay & Wells, 1985). However, in the SEQL it is possible for a foil to be identified prior to the target replacement being observed in a target absent lineup. Given that this error is not included as false positive, the overall result can be an artificial inflation in the SEQL’s capacity to diagnose innocence. This will particularly be the case when target replacements are not highly distinctive, or foils are highly similar. The result in both situations is that foils are more likely to be identified. This point was highlighted in study 4, where if Lindsay and Wells’ (1985) definition of a false positive was used, then the SEQL was more diagnostic of innocence, but if any foil identification was treated as a false positive, then the SIML was more diagnostic of innocence. As noted, the practical and research based problems with the SEQL described were observed in the identification results of this study, and both were avoided using ratings.
Order effects were more pronounced in the SEQL than SIML with regards to identifications, and this was expected, as well as consistent with previous research (Clarke & Davey, 2005). This constituted more targets being identified when presented early. There was also a trend for foils to be identified more when the target was presented late. As noted by Clarke and Davey (2005) the SEQL advantage is dependent on the position of the target relative to the NB alternative. If the target is presented prior to the NB alternative, the target is more likely to be identified, and the inverse is also true. In contrast, there were no order effects with regards to ratings in the SEQL procedure. This constitutes a significant improvement on current practice, though does need to be replicated.

As indicated, the use of ratings also allowed for participants to be aware how many photos they would view and that they would view all photos. This is important as it removes the practical issue of hiding from witnesses how many photos they might view. It also removes potential pressure a witness may feel to make an identification as the lineup approaches its end. This was apparent in the current study with the increased identification of foils late in the SEQL in both target present and absent lineups. This finding has important implications for eyewitness evidence, and may represent a marriage of the best of the SEQL procedure with the advantages of ratings.

Evidence for the SEQL procedure being more diagnostic of accuracy than the SIML was explored earlier, with the SEQL outperforming the SIML across all ratings tasks. Providing further support for SEQL ratings representing a powerful combination is the fact that the SIML procedure slightly outperformed the SEQL with regards to identifications. This means that the SEQL benefit was only observed with regards to ratings.

Some of the factors suggested as being responsible for the SEQL outperforming the SIML with regards to ratings would be less relevant in the real world, however, others would likely retain their influence. For instance, the cognitive and attentional factors affecting the completion of the eight lineups in this study, which may have contributed to the SEQL procedure outperforming the SIML with regards to ratings, would be less likely to influence a real witness’s decision in practice. This is because the gravity of the situation would likely over-ride factors pertaining to inattention, and inhibit heuristic processing. They would also be unlikely to view multiple lineups as was the case in this research. However, the contribution of more global processing and the reliance on more ecphoric rather
than perceptual processes (both features of the SEQL in this research) would be more likely to retain their influence in practice.

*The use of an observation stage to benefit SEQL ratings and identifications.*

Another potential criticism of the SEQL identification procedure is that it can lead to overly conservative responding. The witness only views the lineup once, and in hindsight may regret their decision not to identify a particular lineup member. This issue is also apparent with regards to ratings: the witness may wish to change their previous rating after viewing the next face. The problem is that the witness’s baseline for responding is likely to change across the lineup. This could be managed by allowing witnesses to firstly view the lineup in an ‘observation’ stage (where no decisions are made), and then complete ratings during a second lap. The observation stage might entail the lineup being presented simultaneously, sequentially, or cumulatively (where photos appear one at a time but remain once presented). Including an observation stage might encourage more (unwanted) perceptual comparisons between lineup members, however, the benefits may well outweigh the disadvantages. Further research will bear this out.

The thrust of this research has been that similarity provides a better measure of a person’s memory than identity, and that the identification procedure is deeply flawed. However, if the legal system continues to retain the requirement for the witness to make an identification, then this process might at the very least follow a similar hybrid procedure to that outlined above. This hybrid procedure would involve an observation stage where the lineup is presented (e.g. cumulatively), followed by a second lap where the identification occurs. The inclusion of an observation stage has the potential to both reduce false positives and misses by overcoming several problems with the SEQL identification procedure: premature selection resulting in the target not being viewed, order effects driven by uncertainty regarding the number of faces to be viewed, and the shifting criterion (or baseline) for decision-making.

Research exploring second laps in relation to the SEQL identification procedure suggests that the second lap reduces the response criterion, making any identification (correct or incorrect) more likely. Steblay, Dietrich, et al. (2011) found that witness lineup picks increased from the first to second lap, with the second lap producing more errors than correct identifications. Interestingly, they found that witnesses who elected to view a second lap made significantly more errors than witnesses who either chose to stop after one lap, or were required to view two laps.
Horry, Brewer, Weber & Palmer (2015) also found an increase in the selection rate for the second lap of the SEQL. This reflected both increased correct and incorrect identifications. Of those participants who viewed a second lap, roughly 40% changed their response, most often from a non-identification to identification. However, this research is not comparable to the suggestion being made here for an observation lap for two reasons. First, both studies reported involved two active selection laps, rather than an observation stage where no decision was made. There is likely to be an important difference between these two processes, though further research is need to tease this out. Second, the research available pertains to the SEQL identification procedure, not the SEQL ratings procedure.

**Limitations of similarity ratings**

There are likely to be two main criticisms leveled at similarity ratings based on this research, plus several smaller considerations. First, there is no way to determine whether a high rating was indicative of an identification being made, and second, more targets were identified than rated most similar. Less significant considerations are also addressed which the following: the fact that ratings may encourage more feature-based local processes, which are associated with reduced accuracy; that in the same way that ratings preclude the possibility of a positive identification, they exclude conclusive evidence of a suspect’s innocence; and that there is some level of disconnect between similarity processes and identifications.

**Significant Issue 1: Threshold responding and potential strategies for managing it.**

One issue with similarity ratings is that it is possible for a person to be rated most similar with a large discrepancy between the highest and second highest rating, but still not approach the threshold necessary for an identification to have occurred. Put another way, if an identification represents a certain amount of ecphoric similarity summed, then how is it possible to determine when the threshold has been crossed that would trigger an identification? Clearly the aim of this research has been to highlight the problems with eyewitness identifications, and as such it may appear odd to cite this as a significant issue. However, whilst the witness should not make a determination of sameness, it will remain necessary for a jury to do so: such a determination is necessary for the jury to find a person guilty. Thus, it will still be necessary for the jury to determine whether the suspect is the perpetrator based on all the evidence, and useful for them to consider the meaning of a witness’s rating of a suspect. This is a difficult issue, as to try to interpret a
witness’s rating as indicating a positive identification (e.g. crossing the similarity threshold) is to succumb to the flawed identification paradigm that relies on the witness’s determination of sameness. However, to ignore the threshold issue is to fail to address the potential for ratings to be misleading.

There is a difference between a baseline (where a person typically rates within a certain range – e.g. 0-51%) and threshold (what level of rating would result in an identification). For instance, it is possible (albeit unlikely) for a suspect to be rated most similar, with a high rating, and with a large discrepancy between the highest and second highest rating, but still not cross the threshold necessary for an identification to have occurred: e.g. a witness may say that the suspect is the most similar, much more similar than any other lineup member and highly similar, but still not be the perpetrator. This problematic situation could arise from two factors. First, the ratings could be based on perceptual rather than ephoric factors, meaning that the person was rated highly more on the basis of being visually dissimilar to other lineup members, rather than highly similar to the rater’s memory of the target. Thus the rating would be a product of dissimilarity rather than similarity processes, making it inappropriate for use as evidence of a person’s similarity. Second, it could arise from the fact that a suspect may be highly similar to a witness's memory of the target, without being the target. That a witness gives a high rating to a person who is highly similar to their memory of the target could occur consciously or unconsciously. There is no way to fix the latter, and it represents situations where false positive identifications occur. However, if an opportunity is provided to the witness to indicate the former, then this would reduce it’s impact. It should be noted that incorrect identifications (false positives) can equally occur for the above reasons – the problem is not limited to ratings.

Again, the point is not that ratings are trying to achieve identifications by another method, but that the potential exists for ratings to be misleading. One way to manage this is to employ an algorithm approach similar to Brewer’s research group, where some arbitrary threshold point is determined based on group data. However, this approach is problematic, as data would need to be normed on large groups as well as the individual and in relation to the precise lineup in question. Simply, it is impractical and does not adequately account for the highly subjective nature of similarity. Second, the jury could be provided with the lineup and the witness’s similarity ratings. This would allow the jury to judge whether one person (e.g. the one rated most similar) stands out. If the most highly rated person did stand out in the lineup, this might serve as a warning that the rating was more likely
to involve perceptual processes. Third, based on the results of this research it appears that the SEQL procedure better encourages ephoric factors, and that the SIML may be more susceptible to the influence of perceptual factors. If this finding can be replicated, then this presents further reason for the use of the SEQL over the SIML with regards to ratings. Fourth, to determine whether a high rating indicates an above threshold response, a post-rating question could be added: “Is this person likely to be the perpetrator?” Note the question is not whether they are, but whether they are likely to be the perpetrator. This might constitute a compromise between the two difficulties, though it does flirt dangerously with the identification paradigm. A Likert-scale could be included, such as that described above where the witness selects the most appropriate response to the previous question, with the scale ranging from ‘1’ – “Unsure”, to ‘3’ – “Quite Sure”. Finally, the witness could complete a post-rating statement (either alongside or instead of the Likert-scale) that describes their judgment process, and which would then be presented to the jury.

A final approach, that is radically different, involves rejecting the problem on the basis that it continues to cater to an identification mindset. It can be argued that the issue of an internal threshold needing to be met is less relevant in relation to ratings, unless at some level the question is still being asked of the witness “is this the same person?” Even if it was possible to determine the point where an average threshold rating was to occur (e.g. for person X a rating of above Y equates to an identification), it is not clear exactly how the jury would use this information. The most probable outcome is that the jury would revert to accepting the suspect’s identity based on the witness’s testimony. Instead a new way of viewing eyewitness evidence that relies on similarity without asking the witness the unanswerable question regarding identity is encouraged. In practice, this would entail acknowledging the limitations of similarity and relying more heavily on other supporting evidence (such as distance, duration, lighting, familiarity). This is a radical argument, and one that has significant implications for the courtroom. It involves an enormous conceptual shift on the part of police, the judiciary, and legal representatives. On the other side of this argument, even if witnesses provided ratings that were the responsibility of the jury to interpret, difficulties exist. For instance, even if the jury were provided with context for interpreting a witness’s ratings – e.g. that persons with this unique pattern of ratings correctly rated the target most similar in 80% of cases – jurors cannot ascertain whether this particular witness was in the 80% correct or 20% incorrect. This reflects the difficulties of
applying nomothetic data to idiographic contexts (Melton, Petrila, Poythress, & Slobogin, 2007).

**Significant Issue 2: Targets were rated most similar less often than identified.**

Perhaps the most obvious criticism of similarity ratings based on the results of this research is that targets were identified more often than they received the highest rating in the memory task. This effect was largely in relation to the SIML, with the SEQL procedure being accurate at an equivalent rate to identifications. Potential reasons for why identifications were more accurate than SIML ratings were discussed above, and strategies for managing this explored. However, until future research demonstrates the capacity of the recommendations provided to mitigate this issue, it needs to be acknowledged as an unexpected result. Nevertheless, with this limitation in mind, ratings still offered a range of practical, theoretical and legal advantages over identifications. These include the more nuanced picture provided of the witness’s memory, the collection and presentation of eyewitness evidence that better reflects reality and avoids the illusion of precision associated with identification evidence, and returns the role of determining facts (whether the accused is the perpetrator) to the jury. It is suggested that the net benefit provided by ratings well outweighs the disadvantages.

*It is unclear upon what basis similarity ratings are made.*

Having a better understanding of the basis upon which participants’ base their judgments will assist in altering the ratings process to improve its accuracy. For instance, if further research can show that global/local processes do underlie the differences, then priming more global processing with ratings, and adapting lineup functionality to better cater for this will be important. Thus there is the need for a study that replicates this research, but which obtains information from participants regarding the basis for their ratings and identifications.

*Ratings provide no conclusive statement of innocence.*

Another criticism of ratings might be that they provide no conclusive statement of a suspect’s innocence. The lineup cannot be rejected, thus the suspect is denied the opportunity to be vindicated as innocent. However, this argument is problematic for the reasons relating to a witness making a determination of identity. Whether a foil being rated more similar than a suspect can be treated as a rejection of the lineup is a matter for debate. To regard this as a rejection of the lineup is to provide some evidence for the suspect’s innocence, with the added benefit that the illusion of certainty is removed. Furthermore, some measure of the witness’s
perceived dissimilarity between the suspect and their memory of the perpetrator is obtained. It is useful to know that a suspect was rated 4/7 and a foil as 6/7, or that both a suspect and foil were rated 7/7. Alternatively, to treat it as a conclusive rejection might lead in rare instances to a failure to investigate guilty suspects who were not rated most similar. Thus, a tradeoff exists. While not explored in this research, it would be possible to include an option with ratings whereby the witness can reject the lineup on the basis of no persons being similar to the witness’s memory of the perpetrator. This might provide more conclusive evidence of a suspect’s innocence whilst still avoiding the problems associated with providing conclusive evidence of their guilt.

_The relationship between similarity and identity is imperfect._

Studies 1 and 2 suggested that there are several reasons to expect an imperfect relationship between similarity and identity, including low levels of perceived similarity, different processes used to perceive similarity (e.g. feature-based versus global processing), focus on a different features in feature-based processing (e.g. focus on hair versus mouth), and the fact that a disconnect exists in that a person may be visually similar to a target without actually being them. To this, additional factors may disrupt the relationship between similarity and identity such as differing levels of attention, memory interference, and the subjective or relative nature of similarity – two participants perceive similarity differently (though the same two participants would presumably perceive identity differently as well). As described above, memory interference might exist in the form of a longer retention interval between target observation and rating in the similarity condition compared to the identification condition. It might also include disruptions to memory due to seeing other lineup members before or afterwards. However, this criticism applies equally to identifications. There also appears to be a necessary threshold of similarity in order for similarity to index identifications. Thus, it is possible for a person to be rated most similar (e.g. 3/7 compared to the next highest of 2/7), without them necessarily being identified. This raises the question of how jurors would make sense of ratings evidence provided to them. However, it is important to note that this also exemplifies exactly why similarity ratings are important. It is possible that the person rated 3/7 might be identified, in which case the jury are unaware that such a fine distinction was made, and that the identification may be more prone to error than another involving a larger discrepancy between the highest and second highest ratings. One potential way to overcome this issue might be for all witness’s to complete a number of standardised practice lineups (that have been normed). Their
results on these practice lineups could then be used to further inform their actual judgment regarding the suspect. This would help provide some context to a witness’s ratings that would assist in interpreting their actual eyewitness judgment.

A related issue that again both highlights a problem with, and need for, similarity ratings as an alternative to identifications, is that there was likely more dissimilarity between lineup members in the current study compared to previous research based on the small pool of photos. This means that if more similar persons were used, then the likelihood of one person being differentiated from others is reduced. This seems to indicate that differentiation will be less useful in relation to lineups involving great similarity between persons. However, this is exactly the point. If multiple persons are rated as being similar to a participant’s memory of the target, then how can an identification be expected to be made? This reverts to the original question, ‘do similarity processes underlie identifications?’ If yes, then this potential lack of differentiation is a problem – and is diagnostic of a problem and can be utilised as such. This means that low differentiation can be utilised to sound a warning that a witness may be unsure, or is making unusually fine discriminations.

Future directions and applications

One way to better control some of the extraneous factors described immediately above includes the use of a within participant design where the same participants complete both similarity ratings and identifications. This design would remove the inter-participant subjectivity with regards to similarity ratings and allow for a more precise exploration of the relationship between similarity and identity. It is worth noting that the same person may perceive similarity differently at consecutive time points. This might be due to employing a different analytical process, or due to memory interference, or altered attention levels. However, a within participant design will allow for a better determination of the aspects of ratings that are most clearly related to identifications. For example, it might be that high differentiation is the best predictor of accuracy, and that low differentiation is a good predictor of false positives or misses. A within participants design would also allow for more sophisticated analyses. Such a design could answer the question, if an individual rates a lineup member as most similar, what is the likelihood they will also identify them? The present research could only address this question in general terms. It would also be possible to provide probabilistic statements regarding the relationship between discrepancy and accuracy from a within-participant framework – e.g. that a discrepancy of 3 between the highest and second highest ratings (on a 7-point Likert-scale) results in a correct identification in 90% of cases. A within-participant
design would also allow for a precise examination of whether more targets are identified than rated most similar, and what proportion of targets rated most similar were identified. A larger sample size with a more even gender spread would ensure that results are more generalisable to the wider population.

As noted above, another avenue for improved understanding of ratings would involve examining the basis upon which people make ratings. This might occur by simply asking how a determination of similarity was made. This would provide insight into the circumstances in which global or local judgments are made, and which are more accurate and in what contexts. It would also allow for an examination of whether the SEQL and SIML procedures encourage participants to perform ratings or identifications in slightly different ways. This would aid better understanding of what drives the differences between the two procedures. It would also potentially shed light on why correct identification rates were higher than the number of targets rated most similar. For instance, there may be something inherent in the process of identifications that caused this result. Another design might entail priming participants to employ particular processing strategies. This could include priming local versus global processing, or similarity versus dissimilarity processes. A better understanding of characteristics of judgments made based upon similarity versus dissimilarity may help in better recognising each. Being able to distinguish judgments made on the basis of perceptual dissimilarity (e.g. where a person is identified or rated most similar due to being less unlike the target compared to other lineup members) would be valuable for detecting less reliable eyewitness evidence.

This research only explored picture identification evidence – e.g. involving lineup photographs – for unfamiliar faces. Thus, the use of ratings might be extended to live lineups, and even other forms of evidence such as detection of voice and movement, as well as lineups involving more familiar faces. There are no known reasons for expecting ratings to function differently for other forms of evidence, however, this needs to be demonstrated. This research also focused explicitly on the collection of ratings. Another interesting and necessary future focus would be on juror’s experience of utilising ratings in their decision making, and how they make sense of them. Ratings offer two benefits to jurors, first, they are more nuanced and complex, but this may also represent a difficulty in that jurors may require more guidance in incorporating them into their judgments. Second, ratings return responsibility for determining whether the suspect is the perpetrator to the jury, rather than allowing this to be determined by the witness. Juries’ experience of these two benefits, and what problems they encounter in relying on ratings rather
than identifications are necessary before any change to current process can seriously be advocated.

*Applications.*

The drive of this research has been to suggest (and provide evidence for) similarity ratings as an alternative to identifications in eyewitness evidence. The argument was made that for a witness to ‘identity’ a suspect was both unreliable in the context of psychological research, and inappropriate in terms of legal process (e.g. it usurps the role of the jury as fact finder). However, it is important to acknowledge that whilst a witness’s evidence should be captured and presented in terms of similarity, it will still be necessary for the jury to make a determination of the likelihood of the suspect being the perpetrator. This still involves a determination of sameness, but one that has been removed from the witness and returned to the jury. Two things here, first, it may seem counterintuitive to argue that the witness (who observed the perpetrator) cannot make a judgment of sameness, but the jury (who merely have the witness's account) can. But for the witness to make this determination of sameness is for them to act as both witness and jury, and is inconsistent with other aspects of law. Furthermore, unlike the witness, the jury will make the determination taking into account all the available evidence. Second, for any clarification or conclusion regarding an offender’s identity, there needs to be some judgment of ‘sameness’ by the jury. Without this, any determination of whether the suspect is the same person who committed the crime is impossible. Thus, whilst potentially theoretically flawed (determinations of sameness are unreliable), it is practically necessary to address the issue of identity to make a finding of guilt or innocence possible.

At present, the positive identification of a suspect by one witness is sufficient to lead to conviction. This is a matter of concern that was highlighted long ago by the Devlin Inquiry. Using ratings this situation would be remedied in two ways. First, ratings involve removing the decision from the witness and returning it to the jury. Thus the jury would determine whether the suspect was the perpetrator, not the witness. This is an important change, and one that is more consistent with other aspects of law. Second, a problem with positive identification evidence is that it has typically been afforded great weight by juries. The expression of evidence in terms of similarity rather than identity would reduce its power in the minds of jurors. This also assists in returning the decision making to the jury, who make a determination based on a range of relevant factors, rather than a simple binary decision by the witness. It should be acknowledged that even with identification evidence the Crown
will typically rely on other supporting evidence, and that the defence can potentially present alibi evidence indicating the impossibility of their client being at the crime scene. However, the point that ratings remove the illusion of precision in juror’s remains pertinent. It is possible that prosecutors might be resistant to the use of ratings due to fear that defence lawyers will exploit the reduced certainty inherent in ratings, thus undermining eyewitness evidence. For instance, a defence argument might be as follows: “You rated the accused as 6/7 in terms of similarity. This indicates that you were not sure. If you were certain, why didn’t you rate them as 7/7?” The response to this is that it is up to the prosecution to provide the necessary evidence supporting a witness’s account and to inform the jury regarding ratings. This might be providing information regarding the subjective nature of ratings (e.g. baseline ratings differ between individuals; or for example that only 50% of accurate witnesses assign the maximum similarity rating to a target), the discrepancy between the highest and second highest rating, or probabilistic statements of the likelihood of accuracy based on normative data. It is then the jury’s role to determine whether the accused is the perpetrator based on all the available evidence.

Conclusion

There is a temptation to regard identifications from the traditional modernist perspective that sees a direct correlation between the world and our perception of it. The witness is seen as neutrally observing the world and subsequently reporting on this observation in an objective fashion. However, the past psychological research reported here indicates that what is first seen and later recollected is fashioned by a variety of conscious and unconscious processes involving memory and perception that are subjective. Furthermore, the legal process involving identification evidence is inconsistent with most other aspects of law in that it allows the witness, rather than the jury, to determine fact. The current research explored similarity ratings as an alternative to identifications, and preliminary evidence suggested that a relationship exists between similarity and identity, and that ratings offer a number of advantages over identifications. Advantages of ratings included their providing a more nuanced picture of a witness’s memory, a better representation of reality, returning the role of determining fact to the jury, and allowing for the provision of probabilistic statements regarding the likelihood of an accurate decision that take into account the baseline rating and discrepancy between the highest and second highest rating. While the research conducted here transformed ratings into a binary outcome (whether the target was rated most similar or not), this was merely for the purpose of comparison with identification data. In practice, while asking the question
‘was the target was rated most similar?’ will provide one useful indicator, the situation is to be avoided where traditional identifications are simply replaced with a binary highest rating (yes/no) measure. This could restore the identification process under a new guise.

One important implication of this research is that the combination of ratings with the SEQL procedure represents a powerful new ‘hybrid’ procedure for capturing eyewitness evidence. The SEQL outperforming the SIML with regards to ratings may reflect the fact that SEQL ratings incorporated the strengths of both the SIML and SEQL identification procedures whilst removing their limitations. The main criticisms of the SEQL identification procedure are that not all photos are necessarily seen, that it is susceptible to order effects, and that can encourage overly conservative responding (the witness may fail to pick but regret it later). This also means that backloading (participants being unaware of the number of photos being presented) is necessary. Using ratings, participants were aware of how many photos they would view, were able to observe all photos, and no order effects were present with regards to SEQL ratings, despite such effects being present in the equivalent identification condition. The removal of order effects is an important result, as if the position of a suspect within a lineup affects the likelihood of their selection, this poses considerable difficulties for the integrity of process. As such, the hybrid procedure of the SEQL with ratings may present an important advancement in eyewitness testimony processes, though further research is needed. This SEQL ratings procedure could further be improved by including an ‘observation’ lap, followed by a second lap in which ratings are completed. This would overcome the problem of the baseline for decisions changing across the SEQL lineup: a witness might regret rating a foil presented early as 7/7 after viewing the target. The main limitations of ratings observed were that it was not possible to determine whether a rating met the threshold necessary for an identification, and that SIML ratings were not as accurate as identifications.

While the previous and current research examined here has provided practical, theoretical and legal argument for the use of ratings over identifications, should the requirement remain that a witness identify the perpetrator, then the observation lap occurring prior to an identification lap would also potentially overcome a number of problems with the SEQL identification procedure: that not all photos are necessarily seen, susceptibility to order effects, conservative responding, and the changing baseline for decision-making.
Two metaphors were used to describe identifications in this paper, that of a used car salesman, full of superficial charm, and empty promises, and that of an iceberg, where the greater part lies beneath the surfaces, unobservable and unknown. Similarity ratings may provide a means for a more informed mapping of the eyewitness evidence process, and thus contribute to a fairer and more just legal process. Whilst more complex, ratings offer benefits for all participants in the legal system, including the accused (e.g. the errors associated with identifications that can lead to wrongful convictions are minimised), the witness (e.g. the imposition of a binary decision that does not necessarily reflect their experience is avoided), police (e.g. more detailed information regarding a suspect’s similarity to the perpetrator is obtained), and the jury (e.g. the responsibility for determining whether the accused is the perpetrator is given to the jury and not the witness, and the jury is provided with more information upon which to base this decision). There is currently no other published research exploring the use of similarity as an alternative to identity in the eyewitness context. While this paper has indicated great promise for the use of similarity ratings in eyewitness evidence, there is a need for further research to (1) determine whether these findings will be replicated in research with larger samples, (2) extend them to other forms of evidence (e.g. live lineups and/or familiar faces), (3) examine juror’s experience of relying on ratings to determine whether the accused was in fact the perpetrator, and (4) explore ways to overcome limitations of ratings.
Cases Cited


*Dhanhoa v. The Queen* (2003) 217 CLR 1


*Festa v. R* (2001) 208 CLR 593

*Frye v. United States* (1923) 290 F 1013

*Pitkin v. R* (1995) 69 ALJR 612


*Trudgett v. R* (2008) 70 NSWLR 696

*United States v. Wade* (1967) 388 US 218 Supreme Court
References


Evidence Act (1929).
Identity Crisis in Identification Evidence


Appendices

Appendix A:

Example Plain Language Statement: Similarity condition (Study 1)

The aim of this study is to examine the accuracy of people’s age estimates and similarity ratings for photographs of Caucasian and Asian men and women.

If you agree to participate in this study you will be asked to view 10 photographic lineups, each following the same procedure. First you will be shown a photograph of a target person. You will then be shown a photograph of another person and asked to estimate the age of that person. Finally, you will be shown a photographic lineup and your task will be to rate how similar each member of the lineup is to the target person.

Please read the statements below before deciding whether to participate in the study.

I understand that:

- Ethical approval has been granted by the Human Ethics Advisory Group, Faculty of Health, Deakin University.
- Participation in the study will take approximately 30 minutes.
- My data will be stored in a secure location.
- My data will be strictly anonymous and be treated confidentially.
- My data may be used for associated publications, but no identifiable information will be released.
- My participation is entirely voluntary and there are no penalties for non-participation.
- I have the right to withdraw my participation at any time and for any reason. However, I understand that if I withdraw my participation after commencing the study it will not be possible to identify or erase any of the answers I have already provided.

At the conclusion of this study you can obtain a summary of results by emailing the principal researcher at donald.thomson@deakin.edu.au.

Contact inquiries about this research can be made by emailing the principal researcher at donald.thomson@deakin.edu.au.

If you have any complaints about any aspect of this study, the way it is being conducted, or any questions about your rights as a research participant, then you may contact

- The manager, Research Integrity, Deakin University
  - 221 Burwood Highway, Burwood, Victoria 3125
  - Telephone: (03) 9251-7129 Email: research_complaints@deakin.edu.au
  - Please quote project number [HEAG-HB2-2013]

If you are willing to participate in the study please click "Next"
Example Plain Language Statement: Identification condition (Study 2)

This study is an online identification study. The aim of the study is to determine ways to improve identification procedures in the criminal justice system.

If you agree to participate in this study you will be asked to view 10 identification parades, each following the same procedure. First you will be shown a photograph of a target person. In the interval that follows you will be asked to estimate the age of a different person. Finally, you will be shown an identification parade and asked whether you are able to identify the target person.

Please read the statements below before deciding whether to participate in the study.

I understand that:

- Ethical approval has been granted by the Human Ethics Advisory Group, Faculty of Health, Deakin University.
- Participation in the study will take between 20 and 30 minutes.
- My data will be stored in a secure location.
- My data will be strictly anonymous and be treated confidentially.
- My data may be used for associated publications, but no identifiable information will be released.
- My participation is entirely voluntary and there are no penalties for non-participation.
- I have the right to withdraw my participation at any time and for any reason. However, I understand that if I withdraw my participation after commencing the study it will not be possible to identify or erase any of the answers I have already provided.

At the conclusion of this study you can obtain a summary of results by emailing the principal researcher at donald.thomson@deakin.edu.au.

Contact inquiries about this research can be made by emailing the principal researcher at donald.thomson@deakin.edu.au.

If you have any complaints about any aspect of this study, the way it is being conducted, or any questions about your rights as a research participant, then you may contact:

- The manager, Research Integrity, Deakin University
- 221 Burwood Highway, Burwood, Victoria 3125
- Telephone: (03) 9251-7129 Email: research-integrity@deakin.edu.au
- Please quote project number [HEAG-H92-2013]
Appendix B: Graphic Representation of lineup from Study 1 and Study 2

The target present and absent conditions were collapsed. The proportion in which the person was rated most similar or identified is plotted on the vertical axis, and the lineup members are plotted on the horizontal axis, with NP representing a ‘not present’ selection in the identification condition. Blue lines index similarity ratings and red lines index identifications.

[Graph of Lineup 1]

[Graph of Lineup 2]
Appendix C: Graphic representation of lineups from study 3 and Study 4

The target present and absent conditions were collapsed. The lineup procedure (SIML or SEQL) is also collapsed. The proportion in which the person was rated most similar or identified is plotted on the vertical axis, and the lineup members are plotted on the horizontal axis, with ‘9’ representing a ‘not present’ selection in the identification condition. Blue lines index visual task similarity ratings (perceptual), red lines index memory task similarity ratings (ecphoric), and green lines index identifications.
Identity Crisis in Identification Evidence

Lineup 2 - Target Present (6)

% rated most highly/identified

Lineup members 1-8 (9 - 'not present')

*SIM-Vis*

*SIML-Mem*

*ID*

Lineup 3 - Target Absent (3)

% rated most highly/identified

Lineup members 1-8 (9 - 'not present')

*SIM-Vis*

*SIML-Mem*

*ID*
Lineup 4 - Target Absent (6)

Lineup 5 - Target Present (3)
Identity Crisis in Identification Evidence

Lineup 6 - Target Present (6)

- % rated most highly/identified
- Lineup members 1-8 (9 - 'not present')

Lineup 7 - Target Absent (3)

- % rated most highly/identified
- Lineup members 1-8 (9 - 'not present')
Identity Crisis in Identification Evidence

Lineup 8 - Target Absent (6)

% rated most highly/identified

Lineup members 1-8 (9 - 'not present')

SIM-Vis
SIM-Mem
ID
Appendix D: Questions for witnesses and jurors

Questions for the witness:

• Please provide a brief statement regarding the basis for your eyewitness testimony (e.g. on what did you base your ratings?): __________________________________________
  __________________________________________
  __________________________________________

• How long did you observe the perpetrator for? _______ Seconds/Minutes/Hours (please enter a number and circle the relevant unit of time)

• How much attention did you pay to the perpetrator? Please place an X on the scale below:
  Very little attention               Moderate attention               Close attention

• If you paid moderate or close attention to the perpetrator, what reason did you have for doing so? __________________________________________

• Is the perpetrator familiar to you? Yes/No/Unsure ____________________________(please circle the relevant answer, and add any necessary comment)

• If you circled ‘Yes’, how do you know the perpetrator? __________________________

• Estimate how many times you had seen them previously and in what context:
  Number of time seen previously: __________________________________________
  Context: __________________________________________

• What was the distance between you and the perpetrator when you observed them? _______ Metres (please enter a number in the space provided)

• Please describe the time of day that you observed the perpetrator, whether it was indoors or outdoors, and whether the lighting was natural or artificial:
  Time of day: _______AM/PM; Indoors/Outdoors: _______; Natural/Artificial: _______

• Was the area poorly, moderately, or well lit when you observed the perpetrator? Please place an X on the scale below:
  Poorly lit                           Moderately lit                           Well lit

• Please describe your psychological state at the time you observed the perpetrator:

  __________________________________________

  Please place an X on the scale below:
  Calm                          Moderate fear                          Panic

• Please estimate the length of time between when you observed the perpetrator commit the offence, and viewed the lineup: _______Days/Weeks/Months (please enter a number and circle the relevant unit of time)
Additional questions for the jury to consider in determining the reliability of eyewitness evidence:

- Was the target rated most similar?
- If 'No' was the target rated equally similar as another lineup member?
- Was the rating high overall?
- Was there a large discrepancy between the rating of the suspect and the next highest rating?
- Given the baseline rating and discrepancy, what is the likelihood that the person rated most similar is the perpetrator based on normative data?
Appendix E: Example arrays

Example of a SIML similarity rating array

Example of a SIML identification array
Example of a SIML visual condition similarity rating array (target absent)
Appendix F: Example procedure for Study 1 similarity ratings (screenshots)

Thank you for agreeing to participate in the study: In a moment you will be asked to view 10 photographic lineups. The first two practice lineups will familiarise you with the procedure prior to the eight lineups of the main study.

When viewing the photographic lineups, please be aware that a photograph of the target person may or may not be included in the photographs shown. Additionally, if a photograph of the target person is included, his or her appearance and/or clothing may be different from what you observed in the original photograph.

Please provide the following demographic information and click 'Next' when you are ready to begin the first practice photographic lineup.

What is your age?

What is your sex?
- Male
- Female

What is your ethnicity?
- Caucasian
- African
- Asian
- Middle Eastern
- Aboriginal or Torres Strait Islander
- Pacific Islander
- Mixed ethnicity (more than one of the above ethnicities)
- Other ethnicity (please specify)

Were you born in Australia?
- Yes
- No
For each photographic lineup, you will be shown a photograph of a target person for five seconds.

Target Person

You will then be shown a photograph of a different person for two seconds. You will be asked to estimate this person’s age and to indicate how confident you are that your estimation is accurate.
Finally, you will be shown a photographic lineup containing eight photographs of different persons. These photographs will be shown together.

You will be asked to look at each photograph and in your own time indicate how similar each person is to your memory of the target person.

Practice Photographic Lineup

Moving from left to right, please rate how similar each person is to your memory of the target person by entering a number from 0 to 100 in the corresponding box, where 0 represents ‘not at all similar’ and 100 represents ‘extremely similar’.

You have now completed the first practice photographic lineup. Please click ‘Next’ when you are ready to begin the second practice photographic lineup.