The Butcher, the Baker, the Candlestick Maker:

Investigating Facial Recognition for Multiple-Perpetrator Crimes

by

Alicia Nortje

NRTALI003

Submitted to the University of Cape Town

In fulfilment of the requirements for the degree, PhD

(Doctoral Degree in Philosophy)

in the Faculty of Humanities

University of Cape Town

82 472 words

University of Cape Town

July 2018

Supervisor: Professor Colin Tredoux,
Department of Psychology,
University of Cape Town

Co-supervisor: Dr Annelies Vredeveldt,
Department of Law,
Vrije Universiteit Amsterdam

University of Cape Town

The copyright of this thesis vests in the author. No quotation from it or information derived from it is to be published without full acknowledgement of the source. The thesis is to be used for private study or non-commercial research purposes only.

Published by the University of Cape Town (UCT) in terms of the non-exclusive license granted to UCT by the author.

I confirm that the work in this thesis represents my own work, in concept and execution.

Signed: Alicia Nortje, 6 November, 2018

Signed by candidate

Acknowledgements

First, I must sincerely thank the National Research Foundation, and the Faculty of Humanities at the University of Cape Town. They provided me with valuable funding to assist me with my studies and travel opportunities to present my research at international conferences. Funding opportunities such as these are extremely helpful, especially for young researchers. There is no doubt that I would have struggled greatly without this financial assistance.

I would like to thank my supervisors, Professor Colin Tredoux and Dr Annelies Vredeveldt, who suffered through multiple drafts of this thesis and provided so much valuable input. Thank you for responding to late-night e-mails, reading countless drafts, and for your feedback. Annelies, your quick responses and attention to detail were greatly appreciated, most especially when I could no longer see the errors.

Further gratitude must be extended to Professor Tredoux, who during the final months of writing-up even presented papers on my behalf. I am extremely appreciative of the guidance and mentoring that he gave me throughout my PhD and MA degrees, and his ability to see the big picture saved me from many rabbit holes. Colin, your door was always open for a chat, a cry (sorry!), a “quick” question about work, and feedback, and I thank you for that.

I am also grateful for the useful feedback and insight from Captain Kenneth Speed and Brigadier Gérard Labuschagne about real-world police practices. Additionally, this thesis would not have been possible without assistance from the South Africa Police Service (SAPS) who granted me permission to conduct my research project, as well as the police detectives, detective commanders, and station commanders who participated in this project.

Professor Rebecca Milne, Professor Jacques Py, Professor Siegfried Sporer, and Dr Fanny Verkampt gave me valuable critical feedback on my project, and Professor Milne also reviewed my materials for my police survey. Thank you to Professor Andrew Yonelinas for responding to my e-mails about know/remember instructions. Finally, thank you to Rene Meyer who patiently explained the intricacies of his R package, compANN, to me via e-mail.

The EYE research team in the Department of Psychology at UCT played an important role throughout my postgraduate studies. There are so many team members worth thanking, but I must thank Kate Kempen, Cheneal Puljevic, Stacey Hall, Grace Seutloali, Michaela Deglon, Caroline Allen, and Gary Dalton in particular. Furthermore, the Applied Cognitive Science and Experimental Neuropsychology Team (ACSENT) also provided numerous opportunities for critical feedback and support, especially when I was designing my research proposal. I am also thankful for the guidance and feedback which Kate Kempen, Michelle Hoogenhout, and Lobke Minter provided through e-mails, written feedback, whatsapp voice messages, and statistical advice. The drafts of each chapter were produced during our weekly writing sessions.

An enormous thank-you to my family, extended family and friends, who sent through care packages of food and wine. In particular, I must thank my partner, Jean-Paul da Conceicao, who endured so many dog-walks talking about semantic and episodic memory, IAC models, and eyewitness memory that he should receive an honourary degree in Psychology. He was willing to read chapter drafts, review statistical graphs, and act as a soundboard. His constant support, encouragement, and understanding was irreplaceable. An acknowledgements page is remiss without mentioning my parents who supported me unconditionally. They were always happy to change the topic when I was unwilling to respond to the opening question of each phone call: “Dare I ask…”. Thank you!

Abstract

In the United States, 20% of all violent crime is committed by multiple perpetrators. Despite the prevalence of multiple-perpetrator crimes, most published eyewitness research uses a single-perpetrator paradigm: that is, witnesses view a crime committed by a single perpetrator whom they must recognise later. Multiple-perpetrator crimes, however, present with several problems. Police procedure for administering multiple-suspect parades is poorly defined. Furthermore, eyewitnesses must make multiple identifications, and are tasked with a unique memory problem of perpetrator-role assignment. I studied these problems in the following ways: (a) a survey among South African detectives (N = 75) to investigate how multiple-suspect parades are administered in practice; (b) two face recognition experiments where the number of face-attribute pairs was manipulated at encoding to investigate the effect of set size on both item recognition (for attributes and faces), and associative memory performance (i.e., matching identity to role; N = 70, and N = 67); (c) an eyewitness experiment where participants studied a simulated crime committed by up to 10 perpetrators whom they had to recognise later (N = 200); and (d) a set of simulations testing a revised version of the Interactive Activation and Competition network proposed by Burton et al. (1990) as a computational account of the memory difficulties experienced by eyewitnesses to multiple-perpetrator crimes. Overall, the results suggest that associative memory is particularly vulnerable to the negative effects of set size, and that role-players in law and psychology should consider the implications of these difficulties in court and the laboratory.

Contents

Chapter 1 The Many Faces of Eyewitness Memory for Multiple-Perpetrator Crimes 1

Aims 10

Chapter 2 Identification Parades for Crimes Committed by Multiple Perpetrators 15

Frequency of multiple-perpetrator crimes 16
International statistics 16
South African statistics 18
Characteristics of multiple-perpetrator crimes 19
Number of perpetrators 20
Testing eyewitness memory with identification parades 22
History of identification parades in South Africa 24
Identification parade rules and guidelines relevant to multiple-perpetrator crimes 26
Eighteen parade rules, but only one is relevant ... 26
Logistical challenges with administering identification parades containing multiple suspects 32
Aim and rationale 35
Method 36
Design 36
Sample 36
Police stations 36
Detectives 37
Characteristics of sample 37
Ethics 39
Materials 40
Procedure 43
Results 43
Section one: Experience and training 43
Training 43
Experience 44
Testimony 46
Section two: Investigating crimes 47
Specialisation 47
Section three: Building parades 51
Frequency of multiple-suspect parades 51
Rationale for including multiple suspects 52
Size of parade 56
Logistical aspects 58
Supporting information at identification 62
Discussion 65
Limitations 71
Conclusion 72

Chapter 3 Recognition Memory for Multiple Faces 74

Literature review: Human ability to recognise faces 74
Face recognition versus eyewitness memory studies 75
Limits of human memory for visual items 76
Memory for familiar faces 81
Memory limits for unfamiliar faces 82
Testing for associative memory 90
The role of familiarity and recollection in associative memory tasks 99
Conclusion 101

Chapter 4 Testing the Effect of Set Size on Face Recognition and Associative Memory 103

Aim and rationale..103
Face recognition experiment 1: Testing memory for pairs using a cued-matching task..............105
Method..105
Design 105
Sample 106
Materials 106
Faces at encoding 106
Attributes at encoding 107
Face-attributes pairs at encoding 107
Faces at recognition 108
Attributes at recognition 109
Procedure 110
Results 114
Discriminability and criterion 114
Hits and false alarms for faces, attributes, and pairs across set size.......................................116
Mixed linear model for HR 118
Mixed linear model for FAR 125
Confidence..130
Mixed linear model for confidence 131
Discussion 136
Limitations 139
Face recognition experiment 2: Testing memory for pairs using an Old-New task 141
Method..141
Design..141
Sample 142
Materials 142
Faces at encoding 142
Attributes at encoding 142
Attributes at recognition...................................142
Order of old face-attribute pairs 142
Procedure 143
Encoding.. 143
Filler task 143
Recognition.. 144
Results 147
Mixed linear model for d-prime 148
Mixed linear model for criterion................... 154
ROC curves 159
ROC curves for face recognition.......................163
ROC curves for attribute recognition 165
ROC curves for paired face-attributes recognition test 167
Discussion... 169
Limitations... 174
Conclusion... 177

Chapter 5 Eyewitness Memory for Multiple Perpetrators 179

What is known about eyewitness memory for multiple perpetrators? 180
Estimator versus System variables: Definitions and relationship... 181
Research on eyewitness memory for multiple perpetrators 183
Control group of a single perpetrator............ 188
Number of perpetrators at encoding and recognition... 189
Recognition of all perpetrators 191
Encoding context, role assignment, and memory for roles... 192
Other relevant studies: Unpublished manuscripts and matching tasks.. 194
Aim and rationale 201

Method 202
Design 202
Sample 204
Materials 204
Encoding videos 204
Lineups 205
Procedure 208
Encoding 208
Statement 208
Recognition 208
Results 209
Lineup accuracy for the main assailant 210
Lineup accuracy and role accuracy for all assailants 211
Comparing perfect accuracy for lineups across set sizes 213
Overall lineup accuracy 213
Accuracy for overall lineup identification across all five set size groups 215
Accuracy for overall pairing identification across all five set size groups 216
Discussion 219
Limitations 223
Conclusion 227

Chapter 6 A revised IAC model for Face and Person Recognition of Multiple Perpetrators 229

Models of face recognition 229
Four types of face recognition models 229
Bruce and Young model of face recognition 231
Pictorial and structural codes 231
Recognition of target faces 232
Interactive Activation and Competition (IAC) model of face recognition 236
Interactive Activation and Competition models 236
Computational explanation of the IAC update functions 237
Parameters 238
Activation of units. 238
Net input 239
Inhibitory connections and competition 239
Decay 240
Cycles 240
IAC model to explain associative priming 240
IAC for learning unfamiliar faces 244
Semantic versus Episodic memory 246
Aim and rationale 248
Method 248
Structure of the revised IAC models 248
Scripts and packages in R 251
Results 252
Testing the three models against one another 252
Manipulating set size for Model 3 253
Basic architecture of variations of Model 3 253
Five variations of Model 3 253
Assumptions of the architecture 254
Simulating recognition for perpetrator 1 and associated information 258
Plotting recognition for all units 259
Activation of Person Identity Nodes 257
Activation of Crime Nodes 261
Activation of Role Units 263
Differences in activation levels across models 265
Discussion 267

Limitations 269
Conclusion 272

Chapter 7 General Discussion 273

The aims of the current thesis 273
Recommended and actual police procedure for administering parades in South Africa 274
Face recognition literature and face recognition results 278
Eyewitness literature for multiple-perpetrators 280
Implications of Role Recollection for Multiple-Perpetrator Crimes 282
Strict Criterion: More Questions than Answers 283
Lineup Formats: Experimental Method Versus Real-World Practice 284
What is to be Done About Role Recollection? 286
Computational models and insight from the IAC model 287
Looking forward 290

References 292

Appendices 311

Appendix A: SAPS 329 Form 311
Appendix B: Lineup rules listed in Du Toit et al. (1987) 315
Appendix C: Ethics Approval from the University of Cape Town 317
Appendix D: Ethics Approval from the South African Police Services 318
Appendix E: Survey administered to SAPS officers 321
Appendix F: Old and New Attributes 332
Appendix G: Orders and Pairing of Faces and Attributes 333
Appendix H: Number of trials 334
Appendix I: List of Old and New Attributes . 335
Appendix J: Orders of Faces and Facts 336
Appendix K: Remember and Know Instructions 337
Appendix L: Sample Image from Eyewitness Encoding Video 338
Appendix M: Eyewitness Questionnaire 339
Appendix N: Lineup Instructions 340
Appendix O: Difficulties with Calculating Within-Subject Confidence Intervals for Eyewitness Experiment 341
Appendix P: Parameters for Revised IAC model 342

Tables

Table 2.1 Estimated prevalence of multiple-perpetrator crimes in various countries....17

Table 2.2 Frequencies and percentages of multiple-perpetrator and single-perpetrator crimes across types of crimes as reported in the 2013-2014, and 2016-2017 Victims of Crime Survey..19

Table 2.3 Frequencies and percentages of multiple-perpetrator and single-perpetrator crimes across types of crimes as reported in the 2013-2014, and 2017-2017 Victims of Crime Survey..22

Table 2.4 Demographic and professional characteristics of the sample of police detectives 38

Table 2.5 SAPS survey sections and respective question items 41

Table 2.6 Responses to Section One (training, experience, and testimony) of the survey....44

Table 2.7 Proportion of sample who specialise in each of the six different crime categories 48

Table 2.8 Question items and responses concerning the frequency of multiple perpetrator parades....52

Table 2.9 Motivation for including more than one suspect per parade 54

Table 2.10 Logistical aspects of administering parades 60

Table 2.11 Motivations and responses for preferred parade format....62

Table 2.12 Responses to whether eyewitnesses are able to provide additional information following a positive identification 65

Table 3.1 Experimental design utilised by Podd (1990) to control for load at encoding and recognition 84

Table 3.2 Hit rate, false alarm rate, and d’ for the three experimental groups in Podd (1990) 85

Table 3.3 Experimental design utilised by Metzger (2002) to control for Load at Encoding and Recognition....87

Table 3.4 Study characteristics and results for Podd (1990), Metzger (2002), and Lamont et al. (2005) 89

Table 3.5 Results for face-face pairs and face-names pairs as presented reported in Bastin et al., 2010. 94

Table 3.6 Hit rates and false alarm rates for studies that tested for item memory, and associative memory following encoding of face-name pairs 95

Table 3.7 Hit rate and false alarm rate for actions, faces, and pairs, after two delays as reported in Old and Naveh-Benjamin, 2008....98

Table 4.1 Hit Rate and False Alarm Rate across test type and set size 116

Table 4.2 Characteristics, and chi-square tests of the three models considered for Hits 120

Table 4.3 Coefficients for fixed effects and random effects for Model 3 predicting Hits 121

Table 4.4 Pairwise comparisons between HR of each test within each set size 125

Table 4.5 Characteristics, and chi-square tests of the three models that predict False Alarm Rate 126

Table 4.6 Coefficients for fixed effects and random effects for Model 3 that predicts False Alarm Rate 127

Table 4.7 Pairwise comparisons between FAR of each test within each Set Size 129

Table 4.8 Average confidence for the three tests across set size....................................131

Table 4.9 Characteristics, and chi-square tests of the three models considered for confidence 132

Table 4.10 Model coefficients for Model 3 predicting confidence 133

Table 4.11 Pairwise comparisons between mean confidence of each test within each set size. 136

Table 4.12 A comparison of the face recognition results of Face Recognition Experiment 1, Podd (1990), Metzger, 2002, and Lamont et al. (2005)..........................138

Table 4.13 Descriptive statistics for Hit Rate, False Alarm Rate across test type and set size 147

Table 4.14 Descriptive statistics for d' and c across test type and set size 148

Table 4.15 Characteristics and chi-square tests of the three models predicting d' 150

Table 4.16 Model coefficients for Model 3 predicting d'......................................151

Table 4.17 Pairwise comparisons between test type within set size 154

Table 4.18 Characteristics, and chi-square tests of the three models considered for c...155

Table 4.19 Model coefficients for Model 3 predicting c 156

Table 4.20 Pairwise comparisons between c of levels of test type within set size 159

Table 4.21 Characteristics comparisons of each ROC Curve within set size for faces 164

Table 4.22 Characteristics comparisons of each ROC curve within set size for attributes...166

Table 4.23 Characteristics comparisons of each ROC curve within set size for face-attribute pairs 168

Table 4.24 HR and FAR from Face Recognition Experiment 1 and 2...........................171

Table 5.1 Characteristics and results of empirical research investigating eyewitness memory for multiple perpetrator crimes 185

Table 5.2 Hit Rate and False Alarm Rate reported in each of the three unpublished student dissertations 196

Table 5.3 Assignment of Condition A and Condition B of lineup format across set size 203

Table 5.4 A list of unique roles performed in the various set size crime videos 205

Table 5.5 Hit Rate and False Alarm Rate for the main assailant 210

Table 5.6 Descriptive statistics for all lineup responses across target-present, and target-absent parades 212

Table 5.7 Proportion of participants who made only correct lineup decisions 213

Table 5.8 Coefficients for fixed effects and random effects for Model 2 predicting accuracy for TP parades 215

Table 5.9 Frequencies of participants who achieved either correct or incorrect pairing.. 218

Table 5.10 Coefficients for fixed effects and random effects for Model 2 predicting pairing accuracy for TP parades 218

Table 6.1 Activation Levels of Face Unit 1, PIN Unit 1, Crime Unit 1, and Role Unit 1 across three models 253

Table 6.2 Activation levels for Face Recognition Unit (FRU) 1, Person Identity Node (PIN) 1, Crime 1, and Role 1 across five versions of Model 3.. 258

Table 6.3 Activation levels for Face 1, Person Identity Node Unit 1, Crime 1, and Role 1 across five versions of Model 3..... 265

Figures

Figure 2.1. Geographical distribution of the nine police stations where participating police officers were employed...39

Figure 2.2. A boxplot of the exact estimate of the number of parades formed throughout detectives' career...........45

Figure 2.3. A boxplot of the exact estimate of the number of parades formed in the last twelve months............................46

Figure 2.4. Detectives' exact estimates of how many times they have testified in court. 47

Figure 2.5. A distribution of the estimate of the percentage of crimes committed by multiple perpetrators. 49

Figure 2.6. A distribution of the estimate of the number of perpetrators involved in the investigated crimes....................50

Figure 2.7. A boxplot of the reported greatest number of perpetrators who have committed a single crime investigated by each officer 51

Figure 2.8. A distribution of the responses for the greatest number of suspects that officers have placed together in a single parade. 57

Figure 2.9. A distribution of parade formats preferred by police officers 61

Figure 4.1. Two versions of a target face used at encoding for Face Recognition Experiment 1..................................107

Figure 4.2. An example of an old face and new face used in this experiment ..109

Figure 4.3. Illustration of the procedure used in Face Recognition Experiment 1..................................113

Figure 4.4. Mean Hit Rate and False Alarm Rate across Set Sizes (1,2,3,5,10, 15, and 30) for three tests: faces, attributes, and pairs........................117

Figure 4.5. Hit Rate across set size and test type 123

Figure 4.6. Mean HR across set sizes (1,2,3,5,10, 15, and 30) for three tests (faces, attributes, and pairs)........... 124

Figure 4.7. Main effects of set size, and test type on FAR.................................. 128

Figure 4.8. Mean FAR across set sizes (1,2,3,5,10, 15, and 30) for three tests, faces, attributes, and pairs.............. 130

Figure 4.9. Main effects of set size and test type on confidence......................... 134

Figure 4.10. Mean confidence across Set Sizes (1,2,3,5,10, 15, and 30) for three tests, faces, attributes, and pairs..... 135

Figure 4.11. Illustration of the procedure used in Face Recognition Experiment 2. 146

Figure 4.12. Hits and false alarms as a function of set size, for three types of tests... 148

Figure 4.13. Marginal means for d' across the set size and test type 152

Figure 4.14. Mean d' across Set Sizes (4, 6, 8, 12, 24) for three tests, faces, attributes, and pairs 153

Figure 4.15. Marginal means for c across test type and set size. 157

Figure 4.16. Mean c across Set Sizes (4, 6, 8, 12, 24) for three tests, faces, attributes, and pairs.. 158

Figure 4.17. ROC curves for the three tests, faces, attributes, and pairs across Set Sizes 4, 6, 8, 12 and 24................. 162

Figure 5.1. Target-present version of Target 9, who was the main assailant in version B of all the videos.......................... 207

Figure 5.2. Average d' and c for participants across Set Size groups 214

Figure 5.3. Accuracy for lineup identifications and pairing 217

Figure 6.1. A reconstruction of the theoretical model of face recognition proposed by Bruce and Young (1986) 234

Figure 6.2. Revised version of the IAC model adapted from the PDP Handbook (McClelland et al., 1986) 237

Figure 6.3. A demonstration of how units activate within an IAC network. 238

Figure 6.4. The revised IAC model constructed by Burton et al. (1990) to explain semantic priming. 242

Figure 6.5. A revised and simplified version of the IACL model introduced by Burton (1994)................................245

Figure 6.6. A visual representation of the three models constructed in the current experiment 250

Figure 6.7. A representation of the four pools in the reworked IAC model for eyewitness memory for multiple perpetrators 256

Figure 6.8. This figure is a representation of the reworked IAC for multiple perpetrators and demonstrates the ten-perpetrator condition.....................257

Figure 6.9. Activation levels for Person Identity Nodes (PIN) across the five models after 100 cycles.................260

Figure 6.10. Activation levels for Crime nodes across the five models after 100 cycles. 262

Figure 6.11. Activation levels for Role nodes across the five models after 100 cycles. 264

List of Cases

R v Mkize 1932 (1) PH H17 (N)

Rex v Mokoena 1932 OPD 79

Rex v W [1947] 3 All SA 1 (A)

R v Olia 1935 (SA) 213 (T)

S v Hlalikaya and Others 1997 (1) SACR 612 (SE)

S v Litako and Others 2014 (2) SACR 431 (SCA)

S v Mgedezi 1989 (1) SA 687

S v Mkhize & Others [2009] JOL 23859 (KZD)

S v Mohlathe 2000 (2) SACR 530 (SCA)

S v Mthetwa [1972] 3 All SA 568 (A)

S v Oliphant 2015 JDR 0905 (ECG)

S v Opperman And Another 2010 (643/09) ZASCA 83

S v Phetla 2016 JDR 1225 (GP)

S v Thebus and Another 2003 (2) SACR 319 (CC)

Chapter 1

The Many Faces of Eyewitness Memory for Multiple-Perpetrator Crimes

Alan Crotzer served 25 years for a crime that he did not commit.[1] In 1981, five people were at home in Florida when three assailants – one of whom was armed with a shotgun – broke into their home. The assailants tied up three of the five victims and left them; the assailants abducted the two remaining two and drove to a secluded area. Two of the three assailants raped the victims here. The perpetrator with the shotgun raped both victims.

Three photograph parades were formed in the course of the police investigation. No identifications were made from the first parade. The second parade included photographs of Alan Crotzer and Douglas James. The victim who was raped twice identified Douglas James as the perpetrator who raped only one of the victims, and identified Alan Crotzer as the perpetrator who participated in both rapes and wielded the shotgun. The same victim identified Douglas James' brother, Corlenzo James from a third photo parade. Both Corlenzo James and Douglas James pled guilty, whereas Alan Crotzer pled not-guilty. During the trial, all five victims positively identified Alan Crotzer via a dock identification as the perpetrator who wielded the shotgun. All three suspects were found guilty, and Alan Crotzer was sentenced to 130 years in prison.

In 2003, the DNA samples from the rapes were retested, and the results confirmed that Alan Crotzer was not a biological match. Douglas James and Corlenzo James admitted that they had committed the crime with a friend and that Alan Crotzer had not been involved. Alan Crotzer's conviction was overturned, but by then he had already spent twenty-five years in prison.

[1] https://www.innocenceproject.org/cases/alan-crotzer/

Unfortunately, miscarriages of justice are not uncommon and the fallibility of eyewitness memory is well-established. The Innocence Project reports that more than 70% of the 356 DNA exonerations to date include mistaken eyewitness identification and testimony;[2] and in some jurisdictions, U.S. courts are obliged to caution the jury about the shortcomings of eyewitness memory (for example, see Sheehan, 2011; Wells et al., 2000). Researchers have repeatedly demonstrated that eyewitness memory is unreliable in certain situations: for example, eyewitness memory is poor for perpetrators of a different race (e.g. Malpass & Kravitz, 1969; Meissner & Brigham, 2001), perpetrators wielding weapons (e.g. Steblay, 1992), and across long delays (e.g., Deffenbacher, Bornstein, McGorty, & Penrod, 2008). Most advances in psychological research on eyewitness memory, subsequent policies, and police guidelines are based on research that share the following characteristic: Eyewitness memory is normally tested for a crime committed by only one perpetrator.

Until recently, almost all extant research investigated eyewitness memory for single-perpetrator crimes, even though multiple-perpetrator crimes occur frequently worldwide. South African researchers estimate that between 30% and 50% of rape is committed by groups of perpetrators (Artz & Kunisaki, 2003 as cited in Horvath & Kelly, 2009; Swart, Gilchrist, Butchart, Seedat, & Martin, 2000), and U.S. crime statistics report that roughly 20% of violent crime is committed by multiple offenders (Sourcebook of Criminal Justice Statistics, 2008). Despite the prevalence and frequency of multiple-perpetrator crimes, a small percentage of the existing eyewitness memory literature investigates this topic. At the time of writing, only approximately 14 published studies exist in which eyewitness memory for multiple perpetrators was tested. The results from these studies suggest that eyewitness identification

[2] https://www.innocenceproject.org/eyewitness-identification-reform/

worsens as the number of people to remember increases[3] – that is, as the number of perpetrators increases, so recognition performance decreases. The interpretation of the results of the 14 studies are hindered by the different methods used (see Table 5.1). In some studies, for example, (a) target-absent line-ups were omitted (Clifford & Hollin, 1981; Egan, Pittner, & Goldstein, 1977; Schiff, Banka, & de Bordes Galdi, 1986), (b) eyewitness identification of only one assailant was tested (main assailant: Clifford & Hollin, 1981; Megreya & Bindemann, 2011; any one assailant: Egan et al., 1977; Megreya & Burton, 2006; Shepherd, 1983), (c) only three studies tested eyewitness identification of all perpetrators (Hobson & Wilcock, 2011; Schiff et al., 1986; Wells & Pozzulo, 2006) and (d) only three studies manipulated set size at encoding and included a single-perpetrator control group (Clifford & Hollin, 1981; Megreya & Bindemann, 2011; Megreya & Burton, 2006). Finally, while all of these studies used an eyewitness recognition paradigm, where participants were tested using a line-up (or an array of images) or live individuals, not all of these studies used an eyewitness encoding paradigm, such as a complex visual event like a simulated crime that is presented live or as video.

Partial answers to some of the unanswered questions about eyewitness memory for multiple perpetrators can be found within the face recognition literature. One such question is whether eyewitnesses to multiple-perpetrator crimes can recognise all the perpetrators. Despite the obvious methodological differences between testing face recognition and testing eyewitness memory, it is possible to gather some evidence for the upper limits of human memory for faces from the facial recognition literature where it is common practice to use multiple targets at encoding and recognition (Shapiro & Penrod, 1986).

The research inquiry into eyewitness memory for multiple-perpetrator crimes is not limited to testing only facial recognition. Consider the following scenario: When an eyewitness

[3] Later in this thesis, I will refer to this as set size. Set size refers to the number of perpetrators and the number of targets at encoding for an eyewitness and face recognition experiment, respectively.

to a single-perpetrator crime makes a positive identification from an identification parade, their identification implies that (a) they recognise the individual (b) *as the person who committed actions within the crime*. Thus, two steps occur when an eyewitness makes an identification. The first step is to identify who in the parade, if anyone, is familiar. This is a low-threshold test, because perceiving a face as familiar is not evidence of guilt nor recognition (for a review see Yonelinas, 2002). There are many examples in the eyewitness and face recognition literature (as well as anecdotal evidence) where a face is perceived as familiar, but not remembered. In the second step, the sense of familiarity is accompanied by recollection. For an eyewitness, the familiar individual is recognised as the person who committed the crime.

In contrast, consider an alternative scenario: An eyewitness to a three-perpetrator crime makes a positive identification from an identification parade. Unlike the identification made by the witness to the single-perpetrator crime, this identification *does not* imply the role performed by that individual. Instead, recollection of the roles performed by the identified individual must be elicited independently. The test of eyewitness memory now includes a third step, where the eyewitness must recall the associated roles, actions, and words committed during the crime by the identified person. Requiring eyewitnesses to provide substantiating information to an identification is echoed in the South African judgement[4] given by Judges Legodi, Rabie, and Mabuse:

> The often patent honesty, sincerity and confidence of an identifying witness remain, however, a snare to a judicial officer who does not constantly remind himself of the necessity of disputing any danger of error in such evidence. The witness should be asked by what features, marks or indications they identify the person whom they claim to recognise. Questions relating to height, build, complexion, what clothing he was wearing and so on should be put. A bald statement that the accused is the person who committed the crime is not enough. Such a statement unexplained, untested and uninvestigated, leaves the door wide open for possibilities of mistake.

[4] S v Phetla 2016 JDR 1225 (GP)

In the case involving Alan Crotzer, the victim identified him as the assailant who perpetrated both rapes and was armed with the shotgun. The recollection of roles and actions is not unique to multiple-perpetrator crimes since recollection of the crime is part of the standard investigative procedure (e.g., through a statement or an interview). It is, however, only for multiple-perpetrator crimes where the eyewitness must disentangle perpetrator actions within the crime.

Successful disentanglement of perpetrators and their roles has numerous applied benefits for the purposes of (a) investigation, (b) testing the veracity of eyewitness memory, and (c) sentencing. First, if the police know who did what, then they can tailor their investigation to find the described perpetrators and can confirm which of the individuals described in the statement have been arrested, identified, or are missing.

Second, successful disentanglement of perpetrators has serious consequences for the judicial test of the eyewitness' memory and. provides evidence of the veracity and strength of the witness' statement and testimony. For example, the three presiding judges in the appeal case, S v Phetla,[5] questioned whether the victim had had sufficient opportunity to view the three perpetrators who robbed her. Furthermore, the judges doubted whether her statement was sufficiently interrogated, and consequently, whether the strength of her memory of the crime and her identification decisions were adequately tested. In their judgement, the judges wrestled with the individuation of the perpetrators when recounting what happened during the crime. The victim of the crime was accosted by three male assailants at her business premises, who coerced her inside, ordered her to switch off the alarm of her business, forced her onto the ground, tied her up, and robbed her of her jewellery, hunting rifles, and a pistol. In paragraph 20 and 22, the judges state the following:

[5] *Ibid* footnote 4

> [20] Who forced her on the ground? Who of the three people tied her up with wires?... How long did the process take place and in which direction or whom or what was she looking at when she was so tied up with wires?
>
> [22] Throughout the complainant's evidence, she was not specific as to who did what and in what position she was in relation to each of the three assailants or whether she was looking at each one of them and for how long.

Successful disentanglement has a third benefit: consequences for sentencing. It stands to reason that the most severe sentencing should be reserved for the perpetrator who performed the most severe actions within a crime or who bears the most responsibility for the crime, for example, a main assailant versus an accomplice.[6] There are, of course, situations where it may be difficult to determine who among the perpetrators deserves the most severe sentence (or who did what), for example, in a riot or protest. In the South African judicial system, the principle of common purpose[7] exists to assist the courts when it is not possible to individuate the actions performed by members of a group[8] so that the tenet of causality is satisfied: a crime occurred due to a set of actions perpetrated by an individual. [9] The principle of common purpose, however, is a solution to a *judicial* problem of establishing causality, and did not arise from difficulties in eyewitness memory – it is coincidentally a panacea for both scenarios.

Are eyewitnesses to multiple-perpetrator crimes able to successfully remember the roles committed by each perpetrator? The extracts from S v Phetla[10] suggest that judges think

[6] For an example of this, see S v Opperman And Another 2010 (643/09) ZASCA 83. In this case, two brothers were found guilty of child sexual abuse. The presiding judge, J Motimele, found that one of the brothers (who was the first appellant) was the main assailant, and was the mastermind behind the crimes. For this reason, he received a longer sentence than his brother, his accomplice.

[7] There are several requirements that must be satisfied for the principle of common purpose. This is outlined in S v Mgedezi 1989 (1) SA 687.

[8] This principle is not limited to South Africa. Similar legal principles exist in other countries, such as England, Canada, Australia, Scotland, and the U.S., but do not exist in France and Germany. See note 30 of S v Mgedezi 1989 (1) SA 687.

[9] See the headnotes of S v Thebus and Another 2003 (2) SACR 319 (CC)

[10] *Supra* footnote 4

that eyewitnesses can differentiate perpetrators and roles, but is there any confirmatory evidence from psychological research? And if eyewitnesses cannot, then how would the courts test the eyewitness' testimony and identification? In the appeal case, S v Litako and Others (2014), [11] the judges recount the victim's description of the two individuals who robbed her:

> She described one of them as being tall and dark in complexion. She testified that he had been wearing a cap pulled towards his face so that it could not be seen. She described the second as being light in complexion. Asked whether she could identify them, she answered in the affirmative.

The same witness was asked to attend an identification parade, where she confidently identified two parade members. When asked to describe the actions of the identified individuals, and to explain how she made her identification, the following evidence was provided:

> In a statement to the police Ms Sibanda was emphatic that the first person she had identified at the identification parade was the one who had threatened her with a firearm, demanding money, and that the second person she identified at the identification parade was the one who had both a firearm and a knife and that he was the person who had taken the money. The reason provided in her statement to the police for identifying two people at the identification parade is as follows: 'I have pointed the two suspects because I spent plus seven minutes with them.'

Her justification appears reasonable: her memory could be considered reliable because she had encoded the perpetrators for several minutes. Her lineup identifications, however, were incorrect: She had mistakenly identified two innocent individuals (i.e., foils) from the parade.

The judicial excerpts thus far suggest that the courts believe that eyewitnesses to multiple-perpetrator crimes can differentiate between the actions of each perpetrator. It is unknown how this belief originated within the courts, but the legal test of eyewitness memory for multiple-perpetrator crimes may be an extension of what can be expected of eyewitnesses to single-perpetrator crimes. A similar approach appears to be adopted by researchers within the eyewitness literature: The psychological research and guidelines that have influenced

[11] S v Litako and Others 2014 (2) SACR 431 (SCA)

policy provide limited information about scenarios involving multiple perpetrators (Wells et al., 2000). Only one published manuscript directly addresses parade format for multiple suspects,[12] but this paper does not address parade formats for testing eyewitness memory for multiple perpetrators (Wells & Turtle, 1986). Wells and Turtle recommend that identification parades should be limited to one suspect per parade, and should never consist of only suspects. They argue that the police and courts have no way of assessing the strength of an eyewitness' memory when the parade contains only suspects, because all lineup decisions will lead to an investigation. If the parade is limited to only one suspect who is surrounded by known-to-be-innocent individuals, then some eyewitness decisions are known to be wrong (such as identifying a foil) and will not be investigated. Known-to-be-incorrect decisions demonstrate the (un)reliability of the eyewitness' memory and help preserve police resources for pursuing worthwhile investigative leads. Wells and Turtle posit that in situations where more than one suspect is present, and it is not possible to build multiple parades, then a mixed parade – one that contains more than one suspect, but is not limited to only suspects – is acceptable. Their hypothesis and conclusions were driven by Bayesian probabilities with no empirical data, and despite having a widespread impact on research into eyewitness identification, it remains removed from real-world constraints. Furthermore, their argument is limited to a single-perpetrator crime where the police have multiple suspects, rather than a multiple-perpetrator crime with multiple suspects. This is evident in footnote 2 on page 322:

> This [the authors are referring to the all suspect-parade where each lineup member is a suspect] does not refer to a situation in which there were multiple offenders. Instead, the situation is one in which there is one offender but multiple suspects. The situation of multiple offenders is not considered in the current article.

[12] The terms 'suspect' and 'perpetrator' refer to different individuals. The 'perpetrator' committed the crime. The 'suspect' is the individual who the police believe may be responsible for the crime. Lineups contain 'suspects', and not 'perpetrators', because the line-up is used as an investigative tool to test eyewitness memory during the investigation (and before trial and sentencing). For this reason, 'suspects' can be innocent, but 'perpetrators' are not.

To my knowledge, crimes involving multiple offenders is not adequately considered in any police guidelines for parade administration either. Instead, a similar version of the guidelines proposed by Wells and Turtle (1986) is echoed in the formal and informal policing and legal texts in South Africa. The reasoning that Wells and Turtle provide for a single-suspect parade is relevant for both single-perpetrator and multiple-perpetrator crimes; however, is it realistic? In South Africa, the South African Police Services (SAPS) are required to administer live identification parades, which the courts prefer to photographic parades. However, several logistical requirements arise when holding live parades, for example, SAPS must arrange that the lineup members are physically present at a police station with a formal parade room; lineup members must be physically similar to one another and the suspect; lineup members must be dressed similarly; there should be enough SAPS officers available to transport the lineup members, conduct the parade and look after the witnesses (estimated to be 18 SAPS officers members in total; personal communication with Captain Speed on 18 September 2015); and the time needed to conduct the parade increases since each suspect is afforded the right to change his/her position and number in the parade between witnesses.

The logistical difficulties accompanying live parades are magnified when more suspects are included. In such situations, the police require more unique foils to build an adequately sized parade (each of whom must also share sufficient physical resemblance with the suspects), the parade room must be large enough to accompany all parade members, and more time is needed to form parades for separate suspects and to allow them to change position and number between witness viewings. In South Africa, very few guidelines exist to assist police officers when administering identification parades for multiple-perpetrator crimes, and, consequently, there are few safeguards against poor police practices. Results from Chapter 2 will demonstrate that the existing guidelines are unrealistic and unclear, resulting in police officers adapting their procedure. These results highlight the need to revise current guidelines for administering parades for multiple-perpetrator crimes, especially in South Africa. However, before this can be done, research must demonstrate whether eyewitnesses to multiple-perpetrator crimes are able to recognise all the perpetrators, and identify their roles.

Aims

The aim of the current thesis is to investigate eyewitness memory for multiple perpetrators. The investigation is not limited to face recognition memory, but also examines the memory association between perpetrators and roles of the crime, which is unique to eyewitness memory for multiple perpetrators. Despite its importance, there is a dearth of research investigating this association. This thesis aims to answer two primary research questions: First, are eyewitnesses to multiple-perpetrator crimes able to recognise all the perpetrators responsible for the crime? Second, are eyewitnesses to multiple-perpetrator crimes able to accurately pair perpetrators with their respective roles? A multi-faceted approach is necessary in response to these research questions.

One of the most notable difficulties with administering multiple-suspect parades is implementing the recommendation from Wells and Turtle (1986). This practical challenge poses interesting questions: how police officers navigate such a logistically difficult operation, and have they developed other methods to administer multiple-suspect parades? Secondly, are eyewitnesses in fact required to substantiate their lineup identifications by recalling the actions of the perpetrators? To answer these questions, I surveyed 75 police detectives in the Western Cape, South Africa to better understand how they conduct identification parades (see Chapter 2). Results from the survey showed that detectives often placed two or more suspects (and sometimes all suspects) in one parade. Furthermore, the survey confirmed that eyewitnesses were required to provide supporting information for their identifications; however, it was not clear whether eyewitnesses provided the correct information.

The multiple-perpetrator problem is, primarily, a question about the upper limits of human memory for faces. Both the eyewitness literature and face recognition research yield limited conclusions, since there is a lack of research that answers this question definitively in the eyewitness literature and there are few studies in the face recognition literature that have manipulated set size. Added to which, there is only one study that investigates face recognition memory and associative memory while manipulating set size (Experiment 2B in Bender, Naveh-Benjamin, Amann, & Raz, 2017), but the results are inconclusive. In Chapter 3, I will introduce and discuss the relevant face recognition literature that has tested memory for large numbers of unfamiliar faces. I will also discuss the relevant literature that investigated item memory and associative memory.

Due to the paucity of relevant research, it is necessary to investigate eyewitness memory for multiple perpetrators using both face recognition and eyewitness experiments. In Chapter 4, I report two face recognition experiments in which the number of face-attributes pairs at encoding were manipulated. The materials and methods are identical in both experiments, the only difference between them is the type of recognition test used. In both experiments, participants are shown a varying number of face-attributes pairs, which they must recognise later. In addition to being tested on the individual parts that constitute the pairs, participants are also tested on the pairs. In the first experiment, memory for pairs is tested with a cued-matching task, which is replaced with an Old-New task in the second experiment.

The results from the first face recognition experiment showed that participants could recognise both faces and attributes, separately, even at high set sizes. Performance was always better for attributes than faces. Accuracy for the pairing task, however, was severely impaired by large set sizes to the point where participants were almost unable to make any correct decisions. Results from the second face recognition experiment demonstrated that the detrimental effect of set size on attributes, faces, and pairs persisted, but recognition performance for pairs improved to the same level as faces.

In Chapter 5 I present an eyewitness experiment that tests eyewitness memory for multiple-perpetrator crimes. Participants are shown a video of a simulated crime, which is committed by a varying number of perpetrators. Unlike most other eyewitness experiments where participants are tested on only one perpetrator, participants are tested on all the perpetrators who committed the crime. If participants make a positive lineup identification, then they must also recall the role that the identified person performed during the crime.

Inspection of the lineup results indicated that participants in the multiple-perpetrator conditions performed worse than participants in the single-perpetrator conditions. Furthermore, participants were less likely to correctly recall the role of the identified perpetrators – even if they had made a correct lineup identification.

The test of associative memory is an important addition to the experiments reported in the thesis. Furthermore, the effect of set size on associative memory is novel, and the results are striking. Memory for faces is negatively affected by an increasing set size, but memory for the face-role pairs (i.e., associative memory) is most vulnerable, implying that eyewitnesses to multiple-perpetrator crimes may not be able to accurately recall the roles for identified perpetrators (even if these identifications are correct).

To explain the results of my experiments, I revised the Interactive Activation and Competition (IAC) network originally proposed by Burton, Bruce & Johnston (1990). Unlike other computational models of face recognition, the IAC network incorporates other information associated with the face to construct person-identity. The revised IAC model is discussed in Chapter 6.

Overall, the simulations of the IAC model suggest that the encoding context is the causal nexus for the detrimental effect of set size on eyewitness memory. When eyewitnesses to a multiple-perpetrator crime see any one perpetrator, then the memory of the perpetrator activates. Ensuing this activation, the memory for the for multiple-perpetrator crimes with multiple suspects activates, which in turn, activates all associated information that the crime is linked to –*including* all perpetrators and their roles. Since all perpetrators and all roles are activated, the eyewitness is unable to differentiate among all the activated perpetrators and roles, resulting in both poor lineup performance and role recollection.

The findings of the simulations are not without criticism. Computational simulations are contingent on several assumptions; for example, that all perpetrators are encoded at equal strength. Moreover, it is unknown how differential activation levels map onto conscious experience: Are low activation levels accompanied by low confidence or a sense of familiarity without definite recollection of the memory? I will discuss some of these uncertainties in Chapter 6.

To return to the original question of how do eyewitnesses of multiple-perpetrator crimes perform at a face recognition task, and are they able to accurately pair roles with perpetrators? The results suggest that eyewitnesses are not able to accurately pair roles with perpetrators, and that their performance will decrease as set size increases. This answer, in turns, leads to more questions, for instance, it is unclear whether this a problem of recognition or encoding. A fuller discussion of unanswered questions and limitations are discussed in Chapter 7.

Finally, it needs to be noted that in terms of format and outline of this thesis the intention is to appeal to the reader, and not the typesetter: Therefore, a combination of APA formatting and legal citation style is preferable so that all judicial cases are cited in footnotes. Additionally, the format of this thesis is reflective of the multi-faceted approach. Even though all five studies in this thesis address aspects of eyewitness memory for multiple perpetrators, they do differ substantially from one another. Each study is rooted within its own literature and method, and for the sake of simplicity a detailed literature review is presented at the beginning of each chapter for that study. Furthermore, the discussion for each chapter will demonstrate how that chapter contributes to the overarching question of eyewitness memory for multiple perpetrators. All materials and data can be found at the following link: https://www.dropbox.com/sh/7of48jonzr4msg1/AADXDgcAPdFslXv_uI3b1u08a?dl=0.

Chapter 2

Identification Parades for Crimes Committed by Multiple Perpetrators

This chapter has three aims. The first aim is to introduce the challenges associated with administering identification parades for multiple perpetrators. To do this, I shall review the relevant national and international guidelines for building and administering identification parades, and underscore the paucity of guidelines for administering parades for multiple suspects.[13] The second aim is to investigate the procedure used by South African police officers when administering parades for multiple-perpetrator crimes – specifically, do South African officers follow the recommended procedure, or do they adapt the procedure? The final aim is to determine whether eyewitnesses are expected to substantiate their lineup identification with auxiliary information, for example, the actions that the perpetrator performed during the commission of the crime. To satisfy the second and third aims, I administered a survey to 75 police detectives in the Western Cape. The findings highlight the difficulties that police officers experience with administering identification parades, and demonstrate how police officers adapt recommended procedures when administering parades. Together, these results stress the need to revise current police guidelines for administering parades to include parades for multiple-perpetrator crimes. Furthermore, the findings suggest that eyewitnesses are expected to justify their identifications, but that they may struggle to do so.

[13] In this chapter, I reserve the term 'multiple suspects' for where the police are investigating crimes committed by multiple perpetrators, where each suspect represents a different perpetrator, rather than when the police are investigating a single-perpetrator crime for which they have multiple suspects. When investigating single-perpetrator crimes, the police may be uncertain which of the suspects is the single individual who committed the crime, whereas when investigating multiple-perpetrator crimes, the group of suspects might be thought to be responsible for the crime. Both investigations may present with similar logistical difficulties when building parades, but they do not present with the same memorial difficulties for eyewitnesses (i.e., individuation of perpetrators). For this reason, I ask the reader to assume that 'multiple-suspect' parades refer to multiple-perpetrator crimes to avoid confusion.

Frequency of Multiple-Perpetrator Crimes

International statistics. Not all crimes are perpetrated by lone offenders (see Table 2.1). Criminal statistics in the USA suggest that approximately 20.5% of known violent crimes in 2008 were committed by multiple perpetrators (Source Book of Criminal Justice Statistics, 2008). Furthermore, following a one-year pilot study on lineup procedures, the authors of the 'Illinois report' (Mecklenburg, 2006) documented that multiple-perpetrators crimes occurred almost 40% more frequently than single-perpetrator crimes during that period, and approximately 60% of lineups during that period were for crimes committed by multiple perpetrators.[14] Of the exonerated cases investigated by the Innocence Project,[15] approximately 17.5% were crimes committed by two or more perpetrators. Multiple-perpetrator crimes are not limited to the USA: In the E.U, between 46% and 70% of crimes against minority groups were committed by multiple perpetrators (European Union Agency for Fundamental Rights [FRA], 2012), and 19% of rapes in the United Kingdom were committed by multiple offenders (Curran & Millie, 2003). Of the sexual assaults reported in Australia in 2004, approximately 23% were committed by multiple perpetrators (Australian Bureau of Statistics, 2004).

[14] It is not clear whether there were more line-ups for multiple-perpetrator crimes, because there were more suspects, or whether multiple-perpetrator crimes were more often perpetrated.

[15] https://www.innocenceproject.org/

Table 2.1
Estimated Prevalence of Multiple-Perpetrator Crimes in Various Countries

Country	Source	Crime type	Estimated prevalence
Australia	Australian Bureau of Statistics, 2004	Sexual assault	23.0%
European Union	European Union Agency for Fundamental Rights, 2012	Crimes against minority groups	46.0% –70.0%
South Africa	Artz & Kunisaki, 2003 (as cited in Horvath & Kelly, 2009)	Rape	50.0%
	Maw, 2012	Rape	30.0%
	Statistics South Africa, 2014	Theft	45.9%
		Robbery	79.7%
		Hijacking	100.0%
		Assault	48.4%
		Sexual Assault	36.4%
		Fraud	14.3%
	Swart et al., 2000	Rape	30.0%
United Kingdom	Curran & Millie, 2003	Rape	19.0%
United States	Franklin, 2004	Rape	10.0-33.0%
	Sourcebook of Criminal Justice Statistics, 2008	Violent Crimes (overall)	20.5%
		Assault	75.8%
		Robbery	22.8%
		Rape	1.4%

South African statistics. There are limited statistics provided by the South African Police Service (SAPS) and Statistics South Africa[16] that formally report the frequency of multiple-perpetrator crimes in South Africa. The prevalence of multiple-perpetrator crimes is not reported in the official Crime Statistics report for 2016 or 2017 (South Africa Police Service, 2016, 2017), but the materials used in the Victims of Crime Survey 2013/2014 (Statistics South Africa, 2014) and Victims of Crime Survey 2016/2017 (Statistics South Africa, 2017) detail the number of male and female perpetrators who committed the crime against the victim. These data are not analysed in the final reports, but are available to download.[17] The data from the two Victims of Crime Surveys suggest a higher prevalence of multiple-perpetrator crimes than that reported internationally: Almost half[18] of the victims reported crimes committed by multiple perpetrators in 2013/2014 and 2016/2017 (see Table 2.2). The high prevalence of crimes is mirrored in other South African research: For example, it is estimated that between 30% and 50% of rapes in South Africa is committed by multiple offenders (Artz & Kunisaki, 2003, as cited in Horvath & Kelly, 2009; Swart et al., 2000), and of the sample of South African respondents included in recent research on the psychological impact of rape, 30% were assaulted and raped by multiple perpetrators (Maw, 2012). Furthermore, of the 1886 police dockets of reported rape in Johannesburg analysed by Jewkes et al. (2012), 17% detailed rapes committed by multiple perpetrators. In a field study investigating whether the Eye-Closure Interview improved eyewitness recollection of the

[16] http://www.statssa.gov.za/

[17] These data can be downloaded from the UCT DataFirst website: https://www.datafirst.uct.ac.za/

[18]A small percentage of participants reported this information in the Victims of Crime Survey 2013/2014 (Statistics South Africa, 2014), and Victims of Crime Survey 2016/2017 (Statistics South Africa, 2017). Respondents were asked to estimate the number of perpetrators who fell within predetermined age brackets. For the 2013/2014 survey and 2016/2017, only 426 (.017%) of the 25 336 participants, and 624 (0.029%) out of 21 323 participants provided this information respectively. It is unclear why so few participants reported this information.

crime, 55% of the eyewitnesses who were interviewed reported crimes committed by two or more perpetrators (Vredeveldt et al., 2015).

Table 2.2

Frequencies and Percentages of Multiple-Perpetrator and Single-Perpetrator Crimes across Types of Crimes as reported in the 2013-2014, and 2016-2017 Victims of Crime Survey

Type of crime	Single-perpetrator	Multiple-perpetrator	Total cases
2013-2014 Victims of Crime Survey			
Assault	51.6%	48.4%	217
Fraud	85.7%	14.3%	14
Hijacking	-	100.0%	5
Robbery	20.3%	79.7%	59
Sexual assault	63.6%	36.4%	22
Theft	54.1%	45.9%	109
Overall average	49.1%	50.9%	426
2016-2017 Victims of Crime Survey			
Assault	52.8%	47.2%	159
Consumer fraud	38.5%	61.5%	13
Hijacking	-	100.0%	2
Home robbery	34.7%	65.3%	75
Housebreaking/robbery	61.7%	38.3%	180
Motor vehicle vandalism	50.0%	50.0%	16
Murder	71.4%	28.6%	7
Property vandalism	69.2%	30.8%	26
Robbery	9.1%	90.9%	55
Sexual offence	88.2%	11.8%	17
Theft of crops	100.0%	-	6
Theft of livestock	63.6%	36.4%	22
Theft of motor vehicle	57.1%	42.9%	7
Theft out of motor vehicle	42.3%	57.7%	26
Theft of personal property	38.5%	61.5%	13
Overall average	50.8%	49.2%	624

Note. Percentages are calculated row-wise. The data reported in this table were analysed from the raw data reported in the two Victims of Crime Survey (Statistics South Africa, 2014, 2017). The types of crimes differed between the two surveys, with the 2016-2017 Victims of Crime Survey listing more categories of crime.

Characteristics of Multiple-Perpetrator Crimes

It is evident from crime statistics that multiple-perpetrator crimes occur frequently, and that certain types of crimes are more often perpetrated by multiple offenders than one offender. South African crime statistics also suggest that multiple-perpetrator crimes differ qualitatively from single-perpetrator crimes. Multiple-perpetrator crimes are typically more violent and are more likely to lead to serious injury to the victim (Statistics South Africa, 2014). While victims are less likely to resist when a robbery is committed by increasing numbers of perpetrators, victims are more likely to suffer serious injury if they do (Statistics South Africa, 2014).

Several South African studies reported that the characteristics of rapes committed by one offender differ from rapes committed by multiple offenders. For example, compared to victims of single-perpetrator rapes, Jewkes et al. (2012) found that significantly more victims of multiple-perpetrator rapes reported that the perpetrator was armed and that they did not know their perpetrators or that they only knew them by sight. Multiple-perpetrator rapes were also more likely to occur when the victim was in a public/open space or outdoors, irrespective of whether they were accompanied (Jewkes et al., 2012). Da Silva, Woodhams, and Harkins (2013) reported that rapes committed by groups of three or more perpetrators were significantly longer than when committed by two perpetrators or one perpetrator (5.5 hours, 3.5 hours, and 2.5 hours, respectively).

Number of Perpetrators

In South Africa, multiple-perpetrator rapes are frequently committed by dyads, that is, groups of two perpetrators (da Silva et al., 2013; Jewkes et al., 2012), although the number of perpetrators varies greatly, sometimes upwards of ten perpetrators. For example, of the police dockets of multiple-perpetrator rapes analysed by Jewkes et al. (2012) 37.2% described rapes committed by three or more perpetrators (up to 17 perpetrators).

Data from the Victims of Crime Survey in South Africa 2013-2014 also shows a wide range of the number of perpetrators involved in various types of crime (see Table 2.3). Hijacking and sexual assault were most frequently committed by dyads, whereas theft, robbery, and assault were most frequently committed by three or more perpetrators[19] – in fact, almost

[19] I excluded the 'Fraud' category of crime from this table, because this category contained only two cases.

15% of assaults were committed by more than 7 and up to 22 perpetrators. The maximum number of perpetrators for hijacking was 4, but was greater than 10 for sexual assault (11), theft (14), assault (22), and robbery (32).

Table 2.3

Frequencies and Percentages of Multiple-Perpetrator and Single-Perpetrator Crimes across Types of Crimes as reported in the 2013-2014, and 2017-2017 Victims of Crime Survey

	Number of perpetrators		
Type of crime	Two	Three or more	Max
2013-2014 Victims of Crime Survey			
Assault	32.4% (34)	67.6% (81)	22
Hijacking	60.0 % (3)	40.0% (2)	4
Robbery	40.4% (19)	59.6% (28)	32
Sexual assault	87.5% (7)	12.5% (1)	11
Theft	40.0% (20)	60% (30)	14
2016-2017 Victims of Crime Survey			
Assault	33.3% (25)	66.7% (50)	14
Consumer fraud	62.5% (5)	37.5% (3)	17
Deliberate damaging/burning of dwellings	12.5% (1)	87.5% (7)	12
Home robbery	53.1% (26)	47.9% (23)	9
Housebreaking/Robbery	56.5% (39)	43.5% (30)	6
Motor vehicle vandalism	50.0% (4)	50.0% (4)	4
Robbery	48.0% (24)	52.0% (26)	10
Theft of livestock	50.0% (4)	50.0% (4)	5
Theft of motor vehicle	-	100.0% (3)	5
Theft out of motor vehicle	33.3% (5)	66.7% (10)	6
Theft of personal property	62.5% (5)	37.5% (3)	17

Note. Frequency of cases within that category is reported in parentheses. Percentages are calculated row-wise. Max refers to the greatest reported number of perpetrators who committed a crime together. The data reported in this table were analysed from the raw data reported in the Victims of Crime Survey, and the types of crimes differed between the two surveys, with the 2016-2017 Victims of Crime Survey listing more categories of crime. Some crime categories were removed because there were too few cases reported.

Testing Eyewitness Memory with Identification Parades

The fallibility of eyewitness memory is well documented. The Innocence Project in the United States estimates that approximately 70% of exonerated cases have involved mistaken eyewitness identification.[20] Furthermore, South African courts have long recognised that eyewitness testimony and identification should be treated with caution, especially if not accompanied by other evidence. [21] Eyewitness testimony and identification, however, is not without merit. While DNA evidence can confirm (or disconfirm) whether the suspect matches the biological material collected from the crime scene, eyewitness memory yields additional information, such as placing the suspect at the crime scene during the event of the crime. Eyewitnesses can also provide evidence about other aspects of the crime that are obviously absent from DNA, for example, the actions, motivations, and speech of the perpetrator/s. The weight of eyewitness memory might be most useful in cases where the victim is the sole witness. For example, in rape cases, consent rather than the act of sexual intercourse is often disputed; DNA evidence can confirm sexual intercourse, but cannot indicate whether the victim gave consent. Thus, eyewitness memory is useful at various stages of the investigation from first reporting the crime to testifying in court, even if it may sometimes be erroneous.

Eyewitness memory serves two purposes: First, eyewitnesses can provide details about the crime, which may initiate an investigation, and second, eyewitness memory can be tested with an identification parade to determine whether the police have the correct suspect. Identification parades can be live[22] (where the lineup members appear in person), a photo-array, or delivered as short video clips of lineup members moving their faces from side to side. Video parades are not used in South Africa, but are used in the United Kingdom with the

[20] https://www.innocenceproject.org/causes/eyewitness-misidentification/

[21] *Rex v Mokoena* 1932 OPD 79; *S v Mthetwa* [1972] 3 All SA 568 (A)

[22] This is known as a ‘formal’ parade in the South African legal literature.

VIPER (Video Identification Parade Electronic Recording[23]) and PROMAT (Profile Matching[24]) systems.

Identification parades have two functions. Police use identification parades as a tool to determine who, if anyone, within the parade is recognised by the eyewitness or victim as the perpetrator. This can confirm whether the police have found the perpetrator. Additionally, identification parades – in theory – act as safeguards against eyewitness misidentification. Unlike dock identifications (Wells et al., 1998) and show-ups (Wells, Leippe, & Ostrom, 1979), innocent suspects are protected from biased circumstances when placed among a reasonable number of physically similar individuals in the parade.

Thus, under reasonable and fair circumstances, identification parades should advance an investigation. However, identification parades should be a fair test of eyewitness memory, but also unbiased towards the suspect. These requirements coupled with multiple examples of mistaken identification motivated the need for a set of lineup guidelines in South Africa (e.g., for a review of the misidentification of Aldoph Beck, see Coates, 1999).

History of Identification Parades in South Africa

The South African legal system is a mixture of Roman-Dutch law, English law, and indigenous law. In South Africa, most legislation pertaining to criminal law originates from English Criminal Law (Joubert et al., 2002). The earliest reported identification parades in South Africa date to 1932[25] and 1935[26] (Rust & Tredoux, 1998), roughly 70 years after the first parade was held in England in 1860 (Davies & Griffiths, 2008; Devlin, 1976).

---

[23] http://www.viper.police.uk/

[24] http://www.promatenvision.co.uk/

[25] *R v Mkize 1932* (1) PH H***17*** *(N)*

[26] *R v Olia* 1935 (SA) 213 (T).

In England, four rules for identification parades were formally outlined in the Home Office Circular of 1925, followed by the Report of the Royal Commission on Police Powers and Procedure (1929, as cited in Davies & Griffiths, 2008). The South African legal system followed the recommendations from the Home Office in England. The original rules stated that (a) the suspect must appear in a parade with other reasonably similar individuals; (b) that the officer investigating the crime may not also administer the parade; (c) that witnesses may not see the suspect, live or in a photograph, before the parade; and (d) all lineup proceedings must be recorded

The judges who presided over the cases of the earliest reported parades in South Africa already expressed concern over the fairness of the parades, suggesting that the courts recognised that there was an accepted procedure for administering identification parades. For example, in Rex v. Olia[27], Judge de Wet alludes to a formal procedure for holding parades by recognising that “the mere fact of adding one man to the parade is certainly not a *proper way* of conducting an identification parade*”* (emphasis added; p. 216).

The rules for identification parades were further refined following the recommendations proposed by Lord Devlin (1976), the judge who was commissioned to investigate miscarriages of justice within the English legal system. The report of his investigation, known as the Devlin report (1976), recommended that eyewitnesses must be given unbiased lineup instructions (warning them that the suspect may not be present), that the parade should be photographed, and that cases which rely solely on eyewitness identification should be treated with caution. Not all of these recommendations were immediately implemented in England, and the latter two recommendations were met with much resistance in particular (Davies & Griffiths, 2008).

[27] *Ibid*

Identification Parade Rules and Guidelines Relevant to Multiple-Perpetrator Crimes

In the South African legal system, the Criminal Procedure Act,[28] published shortly after the Devlin report (1976), allowed for police officers to implement identification procedures.[29] While the Criminal Procedure Act allows for officers to make suspects available for identification, it does not provide any formal guidelines for how to do this; instead, the South African legal system relies on case law for guidance on lineup administration. The case law is summarised in the following two publications, Commentary on the Criminal Procedure Act (Du Toit, de Jager, Paizes, Skeen, & van der Merwe, 1987) and Hiemstra's Criminal Procedure (Kruger, 2017). Furthermore in 2007, SAPS issued the National Instruction 1/2007 (SAPS, 2007) which outlines when and how an identification parade can be held. Kruger (2017) further states that the official form that police officers use to document the details of the parade (SAPS 329), also acts as a guideline for how parades should be built and administered (Appendix A).

Eighteen parade rules, but only one is relevant. The three legal publications that address criminal procedure in South Africa contain similar guidelines about the identification parade (Du Toit et al., 1987; Kruger, 2017; SAPS, 2007). In Du Toit's commentary (see Appendix B), 18 rules were identified to guide fair and proper procedure for conducting identification parades. Of these 18 rules, only one directly addresses a possible multiple-perpetrator scenario. Rule 6 states that a second suspect can be added to the parade so long as the two suspects are physically similar and more foils[30] are added to the parade. The number

[28] Criminal Procedure Act 51 of 1977

[29] This is outlined in Section 37 of Chapter 3 as follows: Any police office may … (b) make a person referred to in paragraph (a) (i) or (ii) or paragraph (a) or (b) of section 36B (i) available or cause such person to be made available for identification in such condition, position or apparel as the police office may determine.

[30] In the psychological literature, the known-to-be-innocent persons who stand alongside the perpetrator are called foils, whereas the police refer to these individuals as bystanders.

of foils required for a single-suspect parade is between 8 and 10, and this increases to between 12 and 16 for a multiple-suspect parade.

The consideration of parade size and the number of suspects who should be placed in a parade together has also been addressed in the psychological literature. With Bayesian probability theory and different conditional probability assumptions, Wells and Turtle (1986) determined that the risk of a false alarm was much higher for parades that contain only suspects (i.e., all-suspect parades) than parades that contain only one suspect (i.e., one-suspect parades). Wells and Turtle further argued that all-suspect parades were of little probative value to police officers, because any identification made by the eyewitness would lead to an investigation. Unlike laboratory experiments where researchers know if the eyewitness made a correct decision, police officers do not know the identity of the perpetrator, and therefore cannot confirm whether a real eyewitness to a crime made a correct identification. It is, however, possible to confirm whether an eyewitness is incorrect. If an eyewitness makes an identification from a parade that comprises one suspect and known-to-be-innocent foils, then police can (a) determine the strength and reliability of the eyewitness memory (especially if the eyewitness identifies a foil, and less so if eyewitness rejects the parade); (b) save resources for leads worth investigating; and (c) confirm (to some extent) whether the suspect was the perpetrator. Thus, Wells and Turtle recommend that police use a single-suspect parade rather than an all-suspect parade. They do concede, however, that it may not be possible to construct multiple one-suspect parades, especially when there are very many suspects. In such situation, they suggest that a mixed parade, which includes multiple suspects and more foils, is a better alternative to the all-suspect parade. In footnote 2, on page 322, Wells and Turtle state the following caveat when describing the all-suspect model:

> This does not refer to a situation in which there were multiple offenders. Instead, the situation is one in which there is one offender, but multiple suspects. The situation of multiple offenders is not considered in the current article.

In a later paper, however, Wells explicitly states that "if there are multiple offenders, each suspect should still appear in his or her own lineup" (Wells, 2006, p 10). The conjecture that all-suspect parades should be avoided is evident in multiple legal texts. Kruger (2017), for example, writes on page 3-6 that "obviously a parade should also not consist of suspects exclusively". Code D of the Police and Criminal Procedure Act 1984 (PACE), which was issued by the Home Office in the United Kingdom (2017), clearly specifies in paragraph nine of Annex B that each parade should consist of only one suspect, although a second suspect can be added so long the two suspects are physically similar and the parade size is increased.[31]

Both PACE, and Rule 6 from Du Toit's commentary address when and how a second suspect can be added to a parade. There is no indication that the other 17 rules identified in Du Toit's commentary should be implemented differently for multiple-perpetrator and single-perpetrator scenarios (Appendix B). There are, however, numerous shortcomings with applying the line-up rules to both scenarios; some of these shortcomings are logistical, whereas others address difficulties with administering the parade. Each shortcoming will be discussed briefly.

One shortcoming is what constitutes a reasonable level of physical similarity between the two suspects. Measuring the optimal level of similarity between two lineup members has most frequently been addressed for foil and suspect, and is well-documented albeit unresolved (e.g., see Clark, 2003; Clark & Godfrey, 2009; Navon, 1992; Tunnicliff & Clark, 2000; Wells, Rydell, & Seelau, 1993). The suspect-foil similarity argument is nuanced, and the degree to which foils should be quantitatively similar to ensure a fair parade is unanswered. For multiple-

[31] The identification parade shall consist of at least eight people (in addition to the suspect) who, so far as possible, resemble the suspect in age, height, general appearance and position in life. Only one suspect shall be included in an identification parade unless there are two suspects of roughly similar appearance, in which case they may be paraded together with at least twelve other people. In no circumstances shall more than two suspects be included in one identification parade and where there are separate identification parades, they shall be made up of different people (paragraph 9, Annexure B, Code D of PACE 1984).

suspect scenarios, measuring physical similarity remains a challenge and it is unclear whether foils in the same parade must match both suspects to the same reasonable extent. In theory, this should be possible, because the rules stipulate that two suspects can be placed in the same parade if they are physically similar, but the physical overlap between the two suspects may not be the same overlap between each suspect and the foils. For example, the two suspects may have long blond hair, but half of the foils have green eyes like the first suspect, whereas the other half have brown eyes like the second suspect. It makes little sense, for example, to construct a parade of nine members where six members only match the first suspect, and the remaining three only match the second suspect – this would comprise two parades of nominal size[32] six and three respectively. The lack of clarity about what constitutes a reasonable likeness between suspects, and subsequently the fairness of the parades, is illustrated in two appeal cases in South Africa. In the first case, [33] the appellant argued that the parade used in the original trial case was not fair. That identification parade contained 15 foils (who were aged between 29 to 40 years) and four suspects (who ranged between 17 to 23 years of age). The judges questioned the height differences among the lineup members: Two of the accused were the tallest members on the parade, and the only foil who matched their height belonged to a different racial group and was much older than both suspects. In the second case, the appellants were originally sentenced to 15 years for housebreaking and robbery with aggravating circumstances.[34] The composition of the parade in the original trial was one of the four pivotal issues addressed by the Judge Eksteen. Both appellants appeared together in a single twelve-person parade, but the second appellant contested that the parade was unfairly biased towards him, because the foils

[32] The size of the parade – including the suspect – is known as the nominal size.

[33] *S v Mohlathe* 2000 (2) SACR 530 (SCA)

[34] *S v Oliphant* 2015 JDR 0905 (ECG)

more closely resembled the first appellant. In the former case, Judge Grosskopf deemed that the parade was not fair towards the appellants (among other comments), whereas in the latter case, Judge Eksteen stated the parade was fair. Both cases demonstrate the subjective nature of reasonable physical similarity, but also the difficulty of determining lineup fairness.

Police are required to conduct a second parade in situations where the two suspects are not physically similar (or are found at different stages during the investigation). As implied in Rule 7 (Du Toit et al., 1987) and stated explicitly in Code D of PACE (1984), the second parade should contain unique foils who did not appear in the first parade. To this end, SAPS must have access to a large enough sample of individuals from the same population group as the suspect. This is already a considerable challenge when building photograph parades and increases in difficulty when parades are conducted live – as required in South Africa.[35]

There are other logistical complexities concerning the administration of multiple-suspect parades. Suspects are allowed to change their position, their clothing, and make other reasonable requests prior to any eyewitnesses viewing the parade (Rule 10, 11, and 12 in Appendix B), but these changes must take place out of sight of any eyewitnesses. These three rules – while ensuring fairness to the suspect – also protract the entire lineup procedure, especially for multiple-perpetrator scenarios. Thus, there is a risk that these rules stop assisting police officers and start obstructing them instead.

The South African courts, however, have recognised that these recommendations should not impede police procedure, and that police officers must be afforded manoeuvrability to perform their duties. In fact, Kruger (2017) clearly states that these rules are guidelines only:

> But the court should guard against putting issues in dispute into too many compartments. One should also guard against making cases involving identity parades too complicated. The police directives for parades, the rules which have crystallized in

[35] While photograph parades are recognised by South African courts, they are considered a suboptimal alternative to live parades that should only be held in situations where live parades are not possible (e.g., when the suspect refuses to participate).

> practice and the opinions expressed, are all guidelines, not laws cast in stone. (Kruger, page 3-7).

Furthermore, the judges who presided over the two appeal cases discussed in the current review both stated that identification parades must be evaluated using 'common sense'. Judge Eksteen[36], for example, ruled that judging the parade was one where "a common sense approach is required", and cited the judgement in S v Mohlathe[37] (2000), where Judge Grosskopf ruled that "Common sense dictates that the non-suspects participating in an identification parade should be similar to the suspect in general appearance". What constitutes common sense in either of two approaches remains unclear.

There are other examples where the logistical difficulties of holding parades are recognised. For example, paragraph 3 of Section 11 of the National Instruction (SAPS, 2007) subtly alludes to an identification parade that may be too large to host, and Judge Eksteen repeats the concern for logistical difficulties linked to holding identification parades:

> Where the parade includes several suspects whose general appearance is markedly different, whether on account of height, build, age or otherwise, care should be taken to ensure that there are sufficient non-suspects whose general appearance approximates that of each of the suspects. *In such circumstances it may be advisable to hold more than one parade, particularly if the number of non-suspects that would be required would result in the parade being unduly large and cumbersome.*" (S v Mohlathe, 2009; emphasis added)

Finally, Kruger (2017) states that the official identification parade form (SAPS 329) has the dual function of documenting the details of parades while also providing guidance about how to conduct identification parades. Instead, the SAPS 329 form appears to contradict the recommended guidelines. Rule 6 of Du Toit's commentary states that parades should be limited to one suspect (or two suspects if they look alike), but the SAPS 329 form includes space for the details of up to four suspects (Appendix A).

[36] *S v Oliphant* 2015, *supra* note 22, at paragraph 15 on page 8

[37] *S v Mohlathe 2000*, *supra* note 21, Headnote

Logistical Challenges with Administering Identification Parades Containing Multiple Suspects

From the current review of the police procedures and recommendations, there is an accepted notion within the legal system that police procedure should not be hindered by unnecessarily complicated rules, that parade rules are guidelines, and that common sense should guide officers. But there is also a danger here: The lack of agreement between recommended guidelines, logistical/practical challenges, and scope for adapting procedures may result in unstandardized police practice among police stations and officers. Unstandardized police practice may be even more likely for scenarios that are unaddressed by the guidelines, such as parades for multiple-perpetrator crimes. For this reason, it is unknown whether police officers follow the recommended guidelines when administering parades for multiple-perpetrator crimes, or if they adapt these guidelines.

There is evidence to suggest that police officers adapt their procedures for multiple-perpetrator parades. Hobson, Wilcock and Valentine (2012) surveyed 29 police forces in the United Kingdom, and asked about the difficulties that officers experience when administering parades for multiple-perpetrator crimes. These police forces do not administer live parades, but instead use Video Identification Parade Electronic Recording[38] (VIPER) or Profile Matching[39] (PROMAT) systems to deliver and administer parades. The parades are delivered digitally and comprise short video clips of lineup members moving their faces from side to side, which are shown sequentially. The video clips of the foils are of volunteers, and form part of a larger database (of more than 60 000 instances) that is curated for constructing these parades.

The accepted procedure for the sequential parade in the England and Wales is that the eyewitness is shown a video of each lineup member one at a time and must respond before

[38] http://www.promatenvision.co.uk/

[39] http://www.viper.police.uk/

proceeding to the subsequent lineup member (Annex A of Code D of PACE, 1986). Each parade is restricted to one suspect, unless there who two suspects who are reasonably similar – in such a scenario, both suspects are placed in the same parade with twelve other people. Eyewitnesses are told that they will view the entire parade twice, and must only make a decision after second viewing. The lineup procedure ends when the eyewitness is ready to make a decision and does not want to view any of the lineup members again.

Hobson et al. (2012) reported that the participating police forces admitted to adapting the administration of multiple-suspect parades due to the difficulties that they had experienced with administering these types of parades. One such difficulty was that eyewitnesses would make all their identifications from a single parade. To prevent eyewitnesses from making all their identifications from a single parade, police officers reported that they adapted the lineup instructions so that eyewitnesses knew that each parade contained only one suspect, and that different parades contained different suspects. In situations where eyewitnesses had made all their identifications from a single parade, some police forces would allow eyewitnesses to continue viewing the subsequent parades. The authors point out, however, that allowing eyewitnesses to continue viewing the parades after making all their identifications serves as negative feedback, since eyewitnesses will realise that they had made a mistake. Finally, police forces reported that eyewitnesses expressed confusion about which suspect they were meant to identify from each parade, and after viewing subsequent parades, eyewitnesses would request to correct a previous choice. Consequently, police officers would allow the eyewitnesses to view all the parades before making any decisions.

The confusion that eyewitnesses experience about which perpetrator to identify is not surprising considering the complexity of a multiple-perpetrator crime. The following excerpt from S v Mkhize and Others[40] (2009) illustrates the complexity of disentangling the identity of each perpetrator in a multiple-perpetrator crime:

> He [the victim] then saw the six males walking past the sides of the motor vehicle and they stood at a distance of approximately 7 meters in front of the car. They had a brief discussion amongst themselves and thereafter confronted them (Maphumulo and the deceased). They split themselves into two groups of three each, the one group proceeding towards the driver's side and the other towards the front passenger side of the motor vehicle. Accused No. 1 was in the group that confronted the deceased. He was in possession of a firearm. As he got to the side of the motor vehicle he tapped onto the side window with the barrel of the firearm and shouted "Open, open, open the window". Immediately thereafter two gunshots rang out, one after the other. The first bullet struck the side window of the driver's door and the other struck the driver's door. Maphumulo then alighted from the motor vehicle and as he did, he was confronted by the three people who were by his side. They demanded money and as they did, they searched him. Accused No. 2 took a wallet containing the bank card and cash in the amount of R120. Accused No.4 took the belt. Accused No. 5 took the deceased's bag from the boot of the car. Inside the bag was a cellular phone belonging to the deceased, a Motorola described in count one. According to Maphumulo, accused Nos 3 and 6 did nothing besides advancing towards him with the group. The accused then fled the scene. At a distance of approximately 20 meters from the scene accused No. 1 fired a shot into the air (S v Mkhize and Others, 2009).

Imagine that the eyewitness to the crime described in the excerpt was asked to view a series of identification parades. It is not clear whether they would know which of the six perpetrators they were meant to identify from each of the parades, if the perpetrators were even present.

Multiple-suspect parades present two additional challenges to eyewitnesses. First, eyewitnesses may be expected to make multiple identifications (one for each suspect). Second, eyewitnesses are required to substantiate their identification by recalling the role that that perpetrator performed. Correctly recalling and matching perpetrator actions to perpetrators is unique to multiple-perpetrator crimes. Unlike eyewitnesses to single-perpetrator crimes whose identification implies that the identified suspect performed all the actions during the commission of the crime, identifications made by eyewitnesses to multiple-perpetrator crimes do not imply which actions were perpetrated by the identified suspect. Instead, role recollection

[40] S v Mkhize & Others [2009] JOL 23859 (KZD)

must be tested explicitly. At this stage, it is unclear whether eyewitnesses to multiple-perpetrator crimes can make multiple identifications successfully, and if they can, are they also able to substantiate their identifications with successful recollection of the actions performed by the identified perpetrator? Unlike a laboratory experiment, police officers will not know whether the eyewitness made a correct perpetrator-action pairing, yet this information would still be asked to substantiate their identification. It is evident that may be a challenging task to complete.

Aim and Rationale

Findings from previous research have demonstrated that police forces adapt their procedures and instructions due to the difficulties surrounding the administration of multiple-suspect parades (Hobson et al., 2012).

In South Africa, a different procedure is used to build and administer parades. The South African Police Services (SAPS) uses simultaneous parades, where all members of the parades are shown at the same time to the eyewitness, whereas in the U.K, a sequential procedure is used (Code D of PACE, 1984; for a review of lineup presentation formats, see Lindsay & Wells, 1985; Meissner, Tredoux, Parker, & MacLin, 2005; Mickes, Flowe, & Wixted, 2012; Wilcock & Kneller, 2011). Furthermore, the South African courts prefer live parades to photo parades.

The lack of formal guidelines for multiple-suspect parades, coupled with the flexibility granted to police to ensure that police procedure is not hindered, can result in situations where unconventional lineup scenarios are implemented in South Africa. Very little is known about how South African police officers conduct identification parades in practice, especially multiple-suspect parades. Although both case law and recommended legal texts yield some insight into police practice, neither demonstrate all the methods currently employed by the police, nor do they provide understanding into why police officers adapt or choose certain practices. Additionally, neither case law nor the legal texts discuss whether eyewitnesses to multiple-perpetrators are expected to state the actions and roles of the identified perpetrator.

To better understand how police officers administer multiple-suspect parades, I conducted a survey among police detectives in the Western Cape. I chose this method so that that I could gain input from the police officers themselves. I had considered reviewing the case files for crimes, but SAPS advised against this, because they would not grant ethical approval for any case information where the suspect was found innocent or was not yet tried. Furthermore, I ran the risk of using a biased sample of cases if I only consulted cases that had gone to court. For this reason, I decided to rather survey the police officers.

The primary aims of the survey were to better understand how police officers conducted multiple-suspect parades, and what types of difficulties, if any, they encountered. The second aim of the survey was to determine whether eyewitnesses were required to justify their identification decisions, and if they were able to do so.

Method

Design

The current study used an exploratory design to better understand the current procedures and methods used by SAPS officers when building identification parades for multiple–perpetrator crimes.

Sample

Police stations. A purposive sampling technique was used to identify the police stations to include in this study. With the assistance of Captain Kenneth Speed, sixteen stations in the Western Cape were identified using the following criteria: (a) an officer from the station frequently requested the assistance of a videographer from the LCRC (Local Criminal Record Centre) to film identification parades, and (b) detectives who had completed an identification parade training course with SAPS and were employed at the station. Ethics permission was granted to contact all sixteen stations. Of the sixteen stations, I was unable to contact the station commanders at four stations, and three stations were unable to assist me during the time frame. Thus, the station commanders at nine police stations agreed to assist me within the data collection time frame.

Detectives. I contacted the station commanders at the participating stations and arranged separate meetings with each station commander to discuss the aims and procedures of the project. After these meetings, I was given the contact details for the detective commander at that respective station. With the assistance of the detective commanders, we arranged that the detectives would complete the survey during the daily detective meeting. I attended the last few minutes of the detective meetings at eight of the participating stations. Of the detectives who were in attendance, the detective commander identified which detectives had experience with administering identification parades and were eligible to complete the survey. At no station were all detectives who attended the daily meeting eligible to participate. The primary reason that detectives who attended the meeting could not participate was that they had never administered an identification parade. This was not an exhaustive method of collecting data: Some detectives were not present at the meetings, because they had to be in court, were on leave, or were on a tracing mission. Most of the questionnaires were completed immediately after the meeting. One of the participating detectives recommended that I contact another detective who was based at a police station that was not on the original list of 16 stations. I did administer the survey to the recommended detective.

Characteristics of sample. In total, 75 detectives participated in the survey, and most were males (84%). Sample characteristics are present in Table 2.4, and a map of the geographical distribution of stations is in Figure 2.1. Most of the sample had a rank of Sergeant (36%), followed by Captain (20%), and then Constable (18%). No other identifying information, such as age or ethnicity, were collected from officers, because officers were reassured that their responses and participation would be kept confidential.

Table 2.4

Demographic and Professional Characteristics of the Sample of Police Detectives

Characteristics	n	Percentage
Station		
Bellville	14	18.67
Elsies River	13	17.33
Illegal Firearms Control and Priority Violent Crime [41]	13	17.33
Kleinvlei	11	14.67
Mfuleni	11	14.67
Wynberg	8	10.67
Bishop Lavis	3	4.00
Kensington	1	1.33
Milnerton	1	1.33
Gender		
Male	63	84.00
Female	12	16.00
Rank		
Sergeant	27	36.00
Captain	15	20.00
Constable	14	18.00
Warrant Officer	9	12.00
Lieutenant Colonel	9	12.00
Lieutenant	-	-
Colonel	1	1.33

[41] This is a subdivision of the Directorate for Priority Violent Crimes (also known as the Hawks).

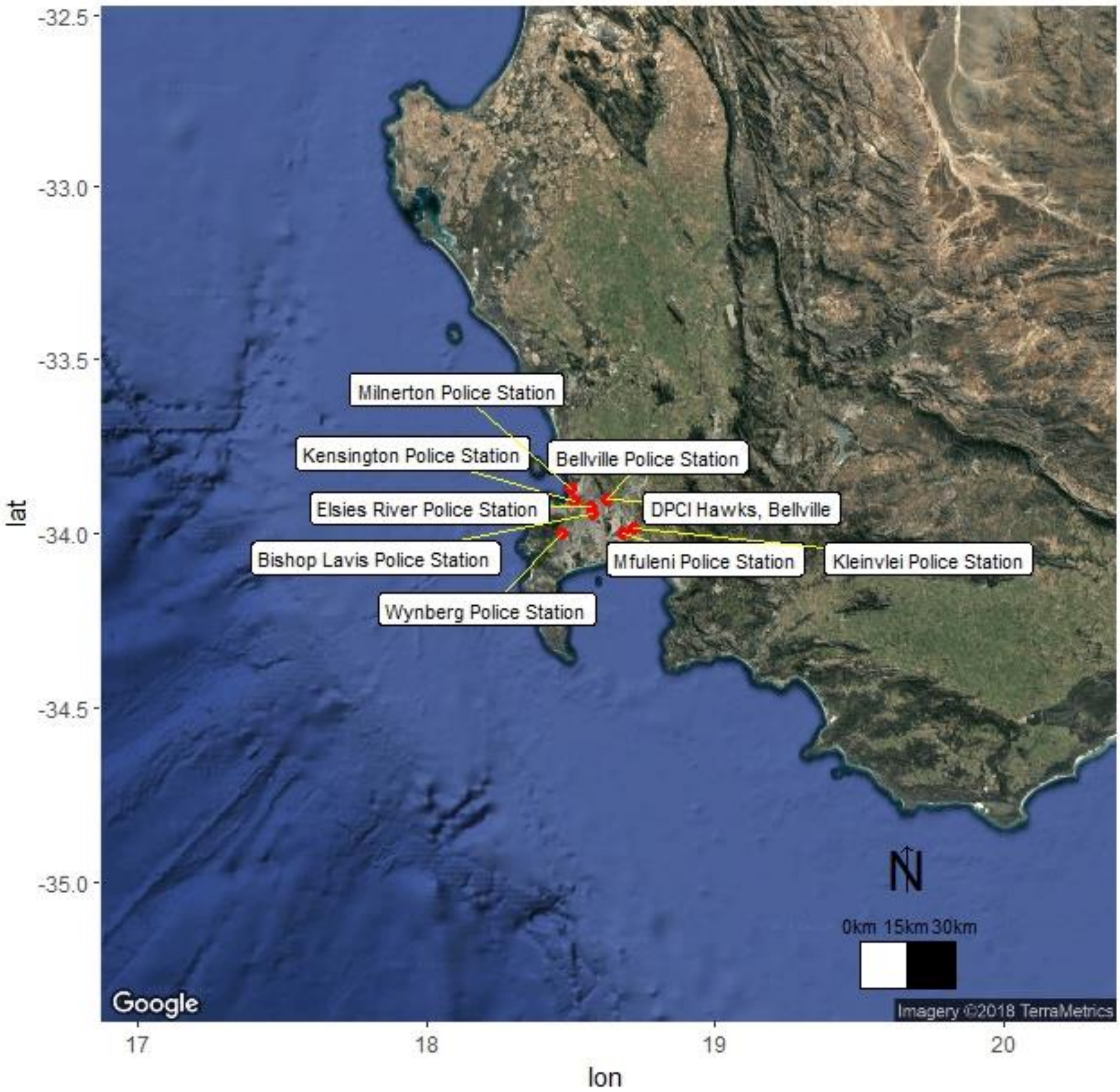

Figure 2.1. Geographical distribution of the nine police stations where participating police officers were employed.

Ethics

Ethics approval was granted by the Ethics Committee of the Department of Psychology at the University of Cape Town (Appendix C). Once approval was granted, ethics approval was granted by the Divisional Commissioner of Research of SAPS (Appendix D). All participants were assured that their responses would be anonymous and confidential, and that their responses would not be used as a measure of job performance nor would their data be made available to anyone besides the researchers. These research conditions were explained to the station commanders and detective commanders of the participating stations.

Materials

The survey for the current study included some of the questions used by Hobson et al. (2012), but was expanded to include additional questions (e.g., specific to live parades). A preliminary version of the survey was administered to two senior detectives at the National Bureau for Illegal Firearms Control and Priority Violent Crime, and was further revised based on their suggestions. The survey was further refined following suggestions from researchers who have experience interviewing and working with the SAPS and police forces in other countries.

The survey was divided into the following three sections, (a) experience and training, (b) investigating crimes, and (c) building parades. Each section comprised multiple questions. The first section included questions about officers' training, rank, and experience giving testimony in court. The second section of the survey asked about the types of crimes that police officers had specialised in. The third section included questions about building and administering identification parades, specifically for multiple-perpetrator crimes. Questions in the third section were grouped into five broad themes, which included frequency and rationale for building multiple-suspect parades, the size of the parade, logistical difficulties for building parades, and whether eyewitnesses were required to provide supporting information at identification. A summary of the questions for each section is listed in Table 2.5, and the complete survey is in Appendix E.

Table 2.5.
SAPS Survey Sections and Respective Question Items

Section	Items
Experience and Training	
Training	1. Did you receive any formal training, including in-service training, about how to conduct an identification parade? 2. What training did you receive about how to conduct an identification parade? 3. What was most helpful about the training 4. What was least helpful about the training, and were there any areas about identification parades neglected by the training? 5. Do you have any recommendations that you think should be added to the training on conducting identification parades?
Experience	1. For how many years have you worked for the South African Police Service? 2. What is your rank in the South African Police Service? 3. *Roughly how many years of experience do you have with conducting / building / administering identification parades? 4. Roughly how many identification parades have you formed/administered across your career? 5. Roughly how many identification parades have you formed/administered in the last 12 months
Testimony	1. Have you ever testified in court about an identification parade that you formed and administered or used in your investigation? 2. Please roughly estimate how many times you been called to court to testify on an identification parade that you formed or used in your investigation across your career
Investigating Crimes	
Specialisation	1. What types of crime do you investigate the most often/specialise in? 2. *Of the crimes that you investigated in the last 12 months, what percentage was committed by multiple (2 or more) perpetrators/criminals? 3. *Of the crimes that you investigate and/or specialise in, how many perpetrators are usually involved/or commit these types of crimes? 4. Of the crimes that you investigate and/or specialise in, what is the greatest number of perpetrators that have committed a single crime? If you specialise in more than one type of crime, please list answers for all of them. 5. *In your experience, which types of crime, including those that you specialise in, are more likely to be committed by multiple perpetrators
Building parades	
Frequency of multiple perpetrator parades	1. Have you ever administered or built an identification parade that contained more than one suspect? 2. Which have you constructed most frequently: parades that contain only <u>one</u> suspect or parades that contain <u>two or more</u> suspects

Rationale for including multiple suspects	1. When, if ever, would you decide to include two or more suspects in one parade? 2. Imagine a situation where a single crime is committed by two or more people. If you are going to build a parade for this crime, and you have two or more suspects who committed this crime together, which method would <u>you prefer</u> to use 3. Please provide reasons for why you prefer the Method that you chose in question 15 (the previous question).
Size of parade	1. Of all the parades held for crimes that you have investigated/administered, what is the greatest number of <u>suspects</u> (for one crime) that you have placed together in the same, single parade? 2. What is a <u>realistic, obtainable</u> number of innocent people (i.e. the other line-up members) to include in a parade 3. What is the largest parade (suspects <u>and</u> other line-up members) that you have ever built/administered?
Logistical aspects	1. How are the <u>other (i.e. innocent people)</u> people who appear in the parade alongside the suspect found? 2. <u>Before</u> the parade is formed, does the suspect have any input in <u>finding/sourcing</u> the people who will appear in the parade with them? 3. Do you more often arrange parades for crimes with multiple *eyewitnesses* (two or more) or single *eyewitnesses*? 4. How much time on average would have passed between when the crime took place and when an identification parade is held? 5. Have you used any of the following parades? 6. If you could choose from the types of parades above, which would you prefer and why?
Supporting information identification	1. When an eyewitness makes identification from the line-up, are they required to state a reason for why/how they made their identification? 2. When an eyewitness makes an identification from the line-up, are they required to state any additional information about the person (e.g. “I recognise him because he had the gun”). 3. In your experience as the officer administering the parade, are eyewitnesses able to name and describe the roles/actions (e.g. “He had the gun”) of all the people whom they identify from the parade? 4. In your experience as the officer administering the parade, do eyewitnesses ever seem confused about the roles and actions of the people whom they identify from the identification parade (that is, they are not certain who did what but they are certain that they were involved)? 5. In your experience <u>as the investigating officer</u>, has it ever happened that the information that the eyewitness provided when making an identification (like in Question 26, 27, 28 and 29) *differed* from the information that they provided in their statement? 6. How did the information that the eyewitness provided when making an identification differ from the information in their statement?

Note. Asterisks indicated questions that were adapted from Hobson et al. (2012).

Procedure

Data collection was arranged with each detective commander, who confirmed the date and time of the meeting and helped identify eligible detectives. The surveys were administered at the end of the detective meeting, and took roughly 35 minutes to complete. Before administering the survey, detectives were told of the purpose of the survey (which was to gain first-hand knowledge of their expertise and work, especially regarding investigations of multiple-perpetrator crimes), were assured that their responses were confidential, anonymous, and would not be used to evaluate their job performance nor professionalism.

Results

As far as possible, the survey contained mainly closed questions, which could be coded quantitatively. The open-ended questions were coded so that they could be analysed quantitatively. A second coder was recruited only for questions where more than 50% of the sample responded, and where the responses were difficult to code. Cohen's kappa was used to check the agreement between the two raters, and on average there was moderate agreement between them, $\bar{\kappa} = 0.61$, 95% CI [0.37, 0.85], $SE = 0.12$; this average is skewed by one agreement rating that was low due to the second coder misunderstanding the coding scheme. Individual Cohen's kappa, with p-values and confidence intervals, will be reported for questions that were coded by both coders. Some of the open-ended questions, especially those towards the end of the survey, had so few responses that a second coder was not needed.

Section One: Experience and Training

Training. Almost the entire sample (90%) had received some form of training about how to conduct identification parades (including building, and administering; see Table 2.6). Of those who received training, most received formal training through either the detective learning programme, resolution of crime course, general training, or an unspecified course.

Table 2.6.

Responses to Section One (Training, Experience, and Testimony) of the Survey

Question Item and Responses	N	%	95% CI
Receive training			
Yes	67	89.33	[82.34, 96.32]
No	8	10.67	[3.68, 17.66]
Type of training			
Detective learning programme	21	27.63	[17.58, 37.68]
Resolution of crime course	15	19.74	[10.79, 28.69]
Specific aspects of training	14	18.42	[9.71, 27.14]
Unspecified course	12	15.79	[7.59, 23.99]
SAPS Paarl college/SAPS academy	6	7.89	[1.83, 13.95]
In-service training	5	6.58	[1.01, 12.15]
Fellow members/colleagues	2	2.63	[0.00, 6.23]
Training from legal experts	1	1.32	[0.00, 3.89]
Number of parades formed throughout career			
0	3	4.00	[0.00, 8.44]
5 or fewer	29	38.67	[27.65, 49.69]
Between 5 and 10	13	17.33	[8.76, 25.90]
Between 10 and 25	7	9.33	[2.75, 15.91]
Between 25 and 50	18	24.00	[14.33, 33.67]
More than 50	5	6.67	[1.02, 12.32]
Number of parades formed in the last 12 months			
0	44	60.27	[49.05, 71.34]
5 or fewer	22	30.14	[19.61, 40.53]
Between 5 and 10	4	5.48	[0.26, 10.63]
Between 10 and 25	2	4.25	[0.00, 8.82]
Between 25 and 50	0	-	-
More than 50	1	1.37	[0.00, 4.00]
Ever Testified in Court			
Yes	34	46.58	[35.14, 58.02]
No	39	53.43	[41.99, 64.87]
Number of times testified in court			
5 or fewer	25	75.76	[61.14, 90.38]
Between 5 and 10	7	21.21	[7.26, 35.16]
Between 10 and 25	1	3.03	[0.00, 8.88]
Between 25 and 50	0	-	-
More than 50	0	-	-

Experience. Participants reported that they served as a functioning member within the SAPS for an average of 19.65 years (*Mdn* = 15.5 years; *SE* = 1.07 years), and had 11.83 years of experience (*Mdn* = 10 years; *SE* = 1.02 years) with identification parades.

Most detectives had formed five or fewer identification parades (38.66%) during their careers, followed by 24% and 17.33% of the sample reporting that they had formed between 5 and 10 parades, and between 10 and 25 parades during their careers respectively. Few participants reported that they had formed 50 or more parades (6.67%), and even fewer reported forming no parades (4.0%).

When prompted for an *exact* estimate of how many parades they had formed throughout their careers, officers reported that they had formed 15.94 identification parades on average (*Mdn* = 5.5; *SD* = 26.76). This average is positively skewed by four extreme responses, with three participants reporting that they had formed at least 100 parades across their careers (Figure 2.2).

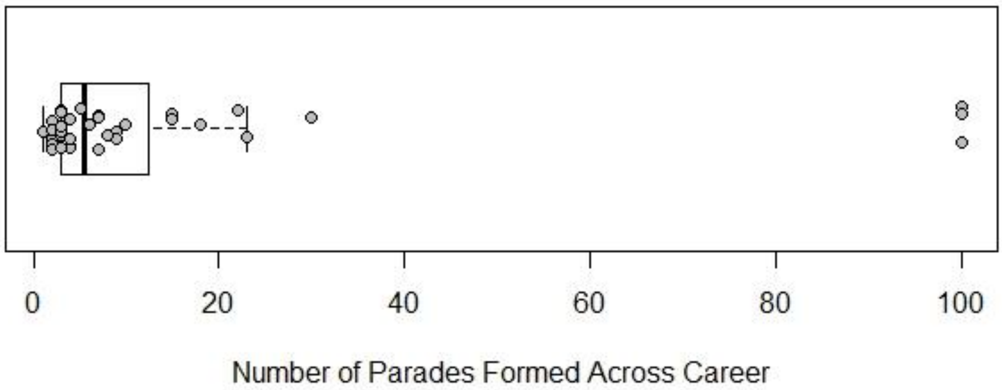

Figure 2.1 A boxplot of the exact estimate of the number of parades formed throughout detectives' career. Individual responses are represented by the grey circles. The median number of parades formed was 5.5, but the mean was 15.94. The boxplot is positively skewed by the four outliers that appear above the right tail of the boxplot.

Most officers reported that they had not formed any identification parades in the last 12 months (60.27%). Of those who had formed a parade in the last 12 months, most had formed five or fewer parades (30.37%).

When asked to provide an exact estimate of how many parades they had formed in the last 12 months, officers reported they had formed three identification parades on average (*Mdn* = 0, *SE* = 1.40). This estimate was skewed by positive outliers (e.g., one participant reported that they had formed 100 parades in the last 12 months; Figure 2.3.).

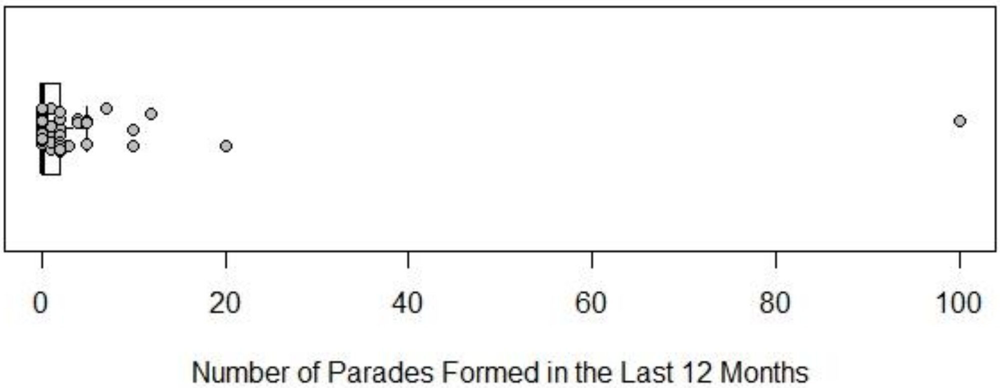

Figure 2.3 A boxplot of the exact estimate of the number of parades formed in the last twelve months. Individual responses are represented by the grey circles. The average number of parades formed in the last twelve months was 3, but the median was 0. The outliers, which appear above the right tail of the boxplot, and especially the officer who reported to have formed 100 parades in the last year, are skewing the average.

Testimony. Two participants did not indicate whether they had testified in court, and their data was removed. Of the remaining participants, 46.58% reported that they had testified in court, and 53.42% reported that they had not. Of the participants who reported that they had testified, most reported that they had testified five or fewer times (75.75%), whereas 21.21% reported that they had testified between five and ten times. On average, officers estimated that they testified in court 3.5 times (*Mdn* = 3, *SE* = 0.59). There was one clear outlier: One officer estimated that they had testified in court 20 times (see Figure 2.4.).

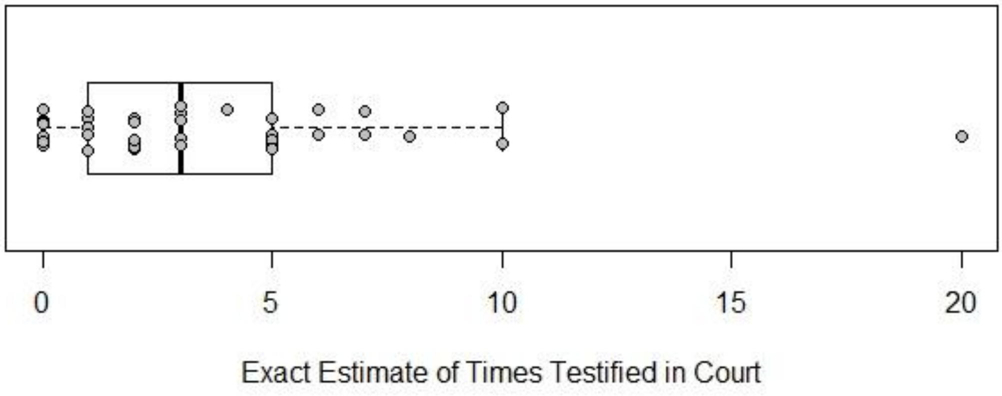

Figure 2.4 Detectives' exact estimates of how many times they have testified in court. Individual responses are are represented by the grey circles. The average number of times that officers have testified is 3.5 (*Mdn* =3, *SE* = 0.59). There is one clear outlier in the boxplot.

Section Two: Investigating Crimes

Specialisation. Participants were asked to list all the crimes that they specialised in, and these responses were grouped according to the crime categories provided in the Crime Statistics Report (SAPS, 2017). Crimes were grouped in the following five broad categories: (a) contact crimes, (b) contact-related crimes, (c) property-related crimes, (d) other serious crimes, and (e) crimes detected due to police activities. I added a sixth category ('Other') for crimes that did not fit the categories listed. The six categories of crimes, and the crimes within each category, are listed in Table 2.7. Since most respondents specialised in more than one type of crime, the percentages reported in Table 2.7 sum to more than 100%.

Overall, most of the participants specialised in contact crimes (78.67%). An equal percentage of the sample specialised in property-related crimes (18.67%) and 'Other' crimes (18.67%). Few participants investigated contact-related crimes (2.67%).

Table 2.7.

Proportion of Sample who Specialise in Each of the Six Different Crime Categories

Category of crime	Examples of crimes from crime statistics report	Other included responses from surveys	%	n	95% CI
Contact crimes	Murder; Sexual offences: rape; Contact sexual offences; Attempted sexual offences; Sexual assault; Attempted murder; Assault grievous bodily harm; Common assault; Common robbery; Robbery aggravated; Carjacking*; Robbery: residential*; Robbery: non–residential*; Robbery: cash in transit; Robbery: bank; Truck hijacking	Serious and violent crimes, ATM bombings	78.67	59	[70.98, 86.36]
Contact-related crimes	Arson; Malicious damage to property		2.67	2	[0.00, 5.70]
Property-related crimes	Burglary at residential premises; Burglary at non-residential premises; Theft of motor vehicle and motor cycle; Theft out of or from motor vehicle; Stock theft		18.67	14	[11.36, 25.99]
Other serious crimes	Other theft; Commercial crime; Shoplifting	Economic crimes, asset and financial crime, fraud, possession of stolen property,	16.00	12	[9.12, 22.88]
Four crimes detected from police action	Illegal possession of firearm and ammunition; Drug related crimes; Driving under the influence of alcohol and drugs; Sexual offences detected as a result of police action	Tracing of suspects	10.67	8	[4.87, 16.47]
Other		Gang cases, general comments about seniority (e.g. “I am the branch commander”), general crimes, inquest, cold cases	18.67	14	[11.35, 25.99]

Note. Asterisk (*) indicates crimes that form part of trio crimes. The types of crimes listed in the ‘Examples of crimes’ column are from the Crime Statistics Report (SAPS, 2017), whereas the types of crimes listed in the ‘Other Included Responses from Surveys’ column are empirical data provided by the survey used in this study. The values in the % column sum to more than 100%, because some respondents specialised in crimes from multiple categories.

Just over half (55.07%) of the officers reported that 50% or more of the crimes that they investigate are committed by multiple perpetrators, and 42.03% of officers estimated that more than 70% of the crimes that they investigate are committed by multiple perpetrators (Figure 2.6).

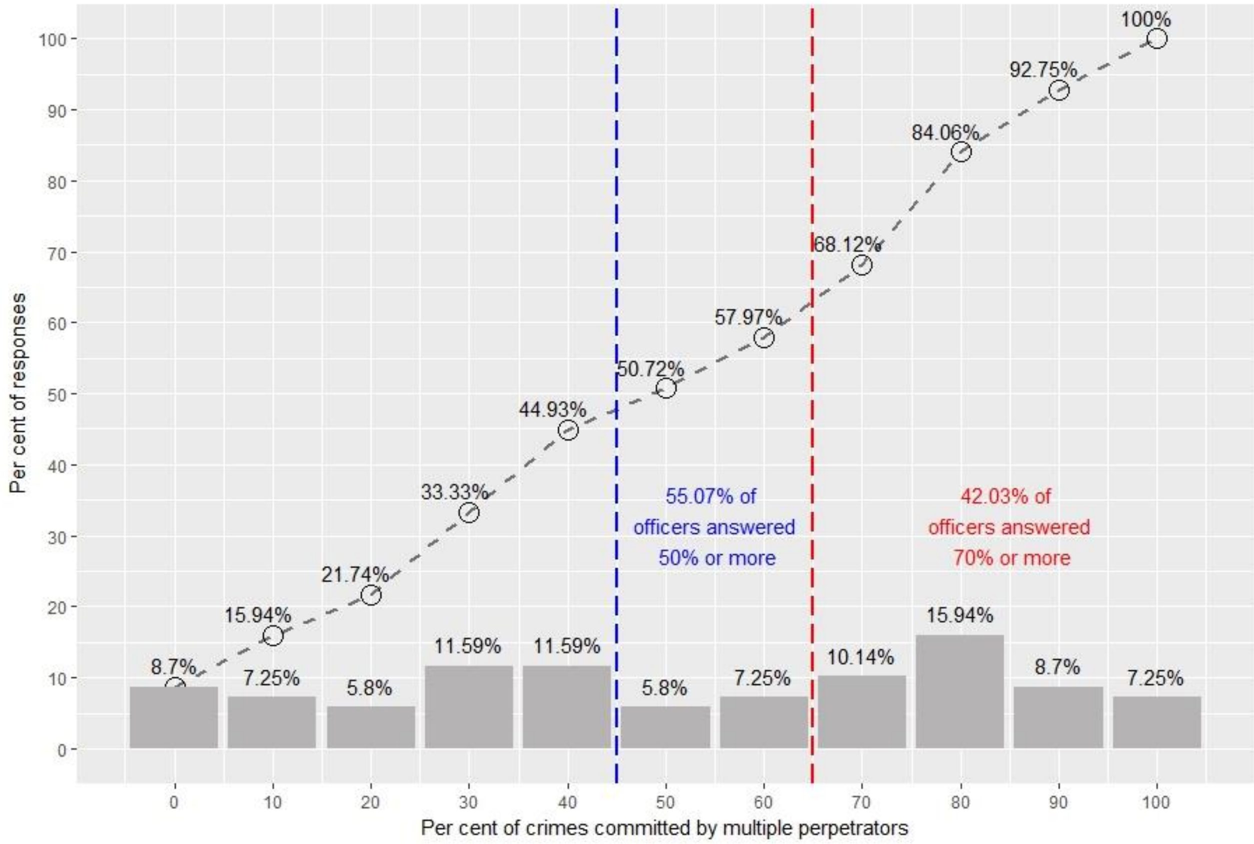

Figure 2.5 A distribution of the estimate of the percentage of crimes committed by multiple perpetrators. The percentage values above each bar represent the response pattern. The dashed line represents cumulative percentage, with these exact values shown next to each circular marker.

Estimates about the percentage of crimes committed by multiple perpetrators (Figure 2.5) were further supported by the detectives' estimates of the number of perpetrators normally involved in the crimes in which they specialised (Figure 2.6). Detectives estimated that 17.59% of the types of crimes that they specialised in were committed by only one perpetrator, whereas the remaining 82.41% of crimes were committed by multiple perpetrators (see Figure 2.6). The most frequent estimate of number of multiple perpetrators involved in a single crime was two

(31.48%), three (20.37%), and four perpetrators (13.89%). Almost 90% of respondents answered that the crimes that they specialised in were committed by up to five perpetrators. Overall, estimates were not limited to small groups of perpetrators, but ranged up to 15 perpetrators who committed a crime together. The types of crimes that were associated with larger estimates (more than five perpetrator) were predominantly cash-in-transit heists and hijackings.

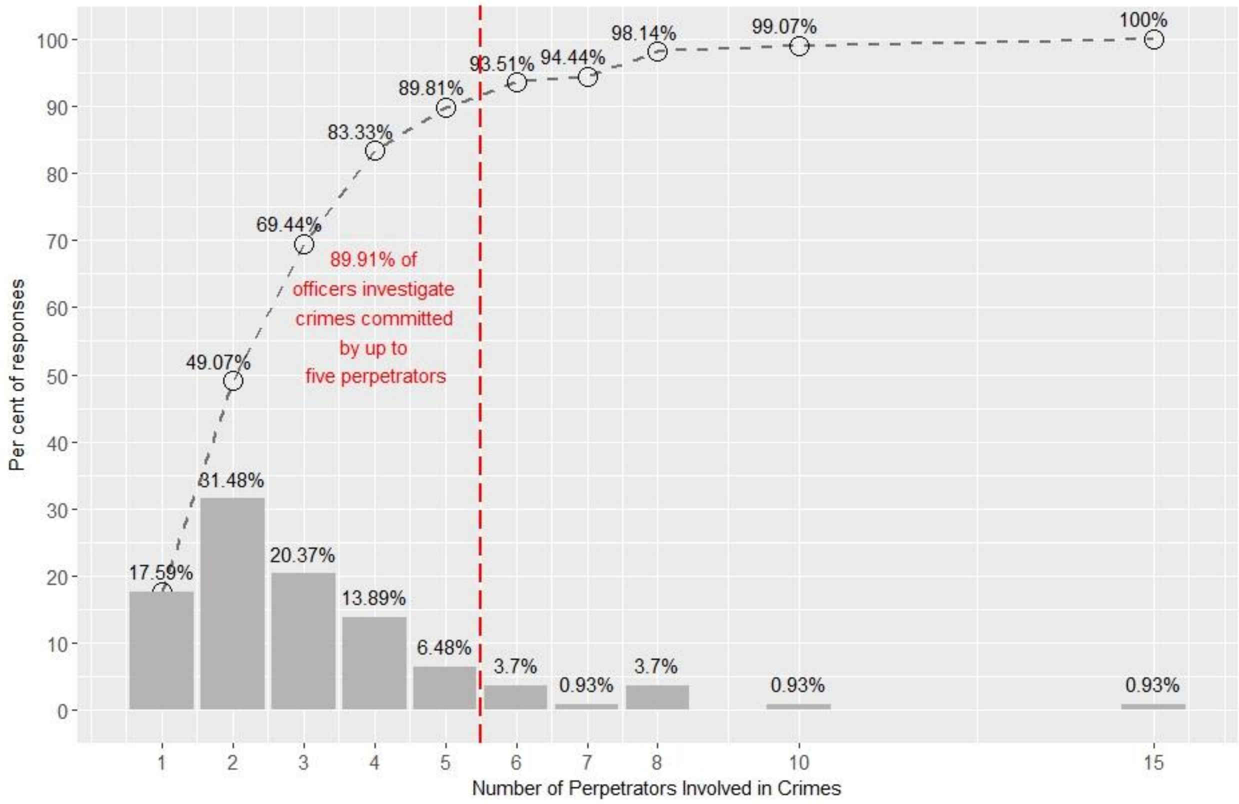

Figure 2.6 A distribution of the estimate of the number of perpetrators involved in the investigated crimes. The percentage values above each bar represent the response pattern. The dashed line represents cumulative percentage, with these exact values shown next to each circular marker.

Of the crimes that they had investigated, police officers were asked to estimate the greatest number of perpetrators who committed a single crime (see Figure 2.7). The most frequently reported numbers were two (13.33%), ten (13.33%), and three perpetrators (10.67%) respectively. On average, the greatest number of perpetrators reported was 7.78 (*Mdn* = 5, *SE*

= 1.01). The greatest number of perpetrators was not limited to 10 perpetrators: 12% of officers reported groups larger than 10 perpetrators,[42] and the largest group reported was 50 perpetrators.

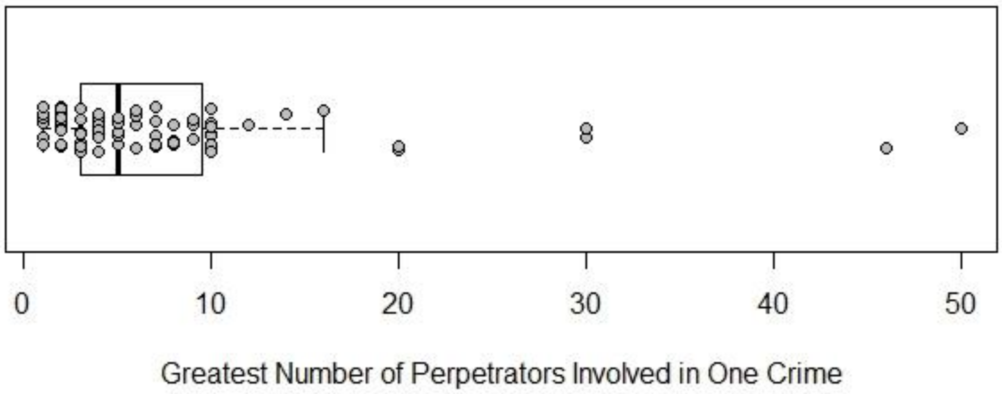

Figure 2.7 A boxplot of the reported greatest number of perpetrators who have committed a single crime investigated by each officer. Individual responses are represented by grey circles.

Section Three: Building Parades

Frequency of multiple-suspect parades. Almost three quarters of the sample reported that they had administered a parade that contained more than one suspect (74.67%; see Table 2.8). The most frequently built parade was one that contained only one suspect (64.38%) rather than two or more suspects (34.25%). One officer indicated that they administered/built both types of parades equally (1.37%).

[42] These values reported here were not limited to just above ten perpetrators. Detectives reported crimes committed by 12, 14, 16, 20, 30, 46, 50, and 500 perpetrators respectively. I was not certain whether the officer who reported 500 perpetrators made a mistake or not. The officer did not indicate the type of crime that was committed by so many individuals (e.g., gang crimes, or drug related crimes). I considered this response as an error and excluded only that response (i.e., 500).

Table 2.8.
Question Items and Responses Concerning the Frequency of Multiple-Perpetrator Parades

Questions	Response Options	n	%	95% CI
Ever administered a parade that contained more than one suspect?	Yes	56	74.67	[64.83, 84.51]
	No	19	25.33	[15.49, 31.17]
Which parade type was constructed most frequently?	Parades with only one suspect	47	64.38	[3.68, 28.32]
	Parades with more than one suspect	25	35.25	[0.29, 21.05]
	Both	1	1.37	[5.57, 31.77]

Note. One officer gave the response 'both' even though it was not listed as a possible option.

Rationale for including multiple suspects. In this section, officers were asked three questions. First, they were asked to explain when (if ever) they would include two or more suspects in a parade. Second, they were asked to imagine a hypothetical scenario where a crime was committed by two or more perpetrators, and that they had arrested two suspects. For the hypothetical scenario, officers had to indicate whether they would build single parade containing both suspects, or two parades containing one suspect each. Finally, officers were asked to motivate their decision for the hypothetical scenario. The results for all three questions are presented in Table 2.9.

For the first question, officers were asked to justify their decision to include (or not include) more than one suspect in the same parade. I reviewed the responses with the assistance of a second coder, and together we generated six themes from the responses given. Following this, we independently coded the responses based on which theme it represented. There was a strong level of agreement between the two raters, $\kappa = 0.78$, 95% CI [0.67, 0.89], $p < .001$. The six themes were split between those who provided reasons for adding multiple suspects, and those who included only one suspect per parade. The data are listed in Table 2.9.

Of the reasons provided by officers to include multiple suspects in a single parade, the most frequently reported reason – which was also surprising – was the number of perpetrators involved or arrested in the crime (74%). Specifically, most officers reported that they would include multiple suspects in a parade if the crime was committed by two or more perpetrators, or if two or more suspects were arrested. Other reasons included the strength of the eyewitness' memory (e.g., "if the complainant says there is more than one suspect that she or she can identify"), to alleviate logistical difficulties (e.g., "always - to save time and because easier to arrange"), or if there were external reasons that allowed for these types of parades to be held (e.g., "if there is enough people to stand parade or the suspects look alike."); these three reasons were reported less frequently (12%, 8% and 6% respectively).

Of the reasons provided for limiting each parade to one suspect, the most frequently reported response was the rules about the parade (e.g., "never allow suspects together in the parade. Have to do each suspect individually") or having never placed more one suspect in a parade together (e.g., "I have never conducted an ID parade with more than one perpetrator"). This reason was reported by 75% of detectives who preferred one suspect per parade. The remaining detectives reported that multiple suspects in a parade could confuse the witness (e.g., "It can confuse the complainant and witnesses").

Table 2.9.
Motivation for Including More than One Suspect per Parade

Questions	**Themes**	**n**	**%**	**95% CI**
Q1: Reasons for parade format				
Parade Format One: Multiple suspects in one parade	1. If more than suspect was involved, or arrested	37	74.00	[61.84, 86.16]
	2. Strength of eyewitness' memory	6	12.00	[2.99, 21.01]
	3. To alleviate logistical difficulties	4	8.00	[0.48, 15.52]
	4. Feasibility based on external factors	3	6.00	[0.00, 12.58]
Parade Format Two: One suspect per parade	1. Has never added more than suspect, or does not allow more than suspect per parade	9	75.00	[50.50, 99.50]
	2. Presents with difficulties for the eyewitness	3	25.00	[0.50, 49.50]
Q2: Hypothetical scenario preference	1. Method One: Build a parade with all the suspects	47	66.20	[55.20, 77.20]
	2. Method Two: Build multiple parades with suspect in each	20	28.17	[17.71, 38.63]
	3. Method Three: Other	4	5.63	[0.27, 10.99]
Q3: Hypothetical scenario motivation				
Parade Format One: Multiple suspects in one parade	1. Better for eyewitness/witness comfort	16	26.23	[15.19, 37.27]
	2. Less time consuming	16	26.23	[15.19, 37.27]
	3. Logistical ease	18	29.51	[18.06, 40.96]
	4. For the courts	5	8.20	[1.32, 15.09]
	5. Most common method/trained that way	3	4.92	[0.00, 10.35]
	6. Other	3	4.92	[0.00, 10.35]
Parade Format Two: One suspect per parade	1. Fair to the suspect	5	25.00	[6.02, 43.98]
	2. For the courts/judicial purposes	5	25.00	[6.02, 43.98]
	3. Trained that way/guidelines	1	5.0	[0.00, 14.55]
	4. Easier for witness	8	40.00	[18.53, 61.47]
	5. Logistical/practical difficulties	1	5.00	[0.00, 14.55]
Parade Format Three: Photographic	Other	4	100.0	-

Following this, officers were presented with a hypothetical scenario where a crime was committed by two or more perpetrators and that police had arrested two or more suspects who may have committed the crime together. For the hypothetical scenario, the respondents had to decide whether they would place both suspects in the same parade, or in two separate parades. Most officers reported that they would place both suspects together in a single parade (66.20%), whereas 28.17% reported that they would place suspects in separate parades (see Table 2.9). Four officers (5.63%) opted for a third parade method, and upon prompting, stated that they preferred a photograph parade.

Officers were asked to motivate their decision for the hypothetical scenario. The same two coders identified the themes from these responses. Six themes were identified for the multiple-suspects-per-parade scenario, and five themes were identified for the one-suspect-per-parade scenario. There was less agreement between the two coders for this question, $\kappa = 0.42$, 95% CI [0.29, 0.54], $p < 0.001$, but this was due to the second coder accidentally coding the one-suspect-per-parade scenario with the six themes identified for the multiple-suspects-per-parade scenario. Overall, for the multiple-suspects-per-parade scenario, officers most frequently reported that this type of parade was easier to arrange and had fewer practical/logistical challenges (29.51%). This category included responses that detailed any practical elements (e.g., easier to arrange, lower risk of something going wrong, easier to administer), but excluded the amount of time that it took to administer, which was a separate category. The second most frequently reported motivation was that a single parade benefitted the witness (26.23%). We coded this category as ‘witness comfort’ as most responses described the emotional wellbeing of the witness. Officers believed that viewing one parade was less traumatic for the witness. The third most frequently reported motivation was that a single parade was less time-consuming (26.23%).

Of the reasons provided for the one-suspect-per-parade scenario, most officers thought that this was easier for the witness (40.0%), because they only had to make one decision per parade. This contrasts with multiple-suspects-per-parade scenario where witness comfort was a common theme. Interestingly, officers also mentioned that this parade type was fairer to the suspect (25%). Finally, officers said that this type of parade was better for judicial purposes (25%). Officers justified this response by referring to the accepted practice recognised by the courts, and that it was easy for the police to demonstrate that the parade was administered fairly.

Size of parade. When asked to estimate the largest number of suspects that police officers had placed in a single parade, 9.23% reported that they opted for a single suspect per parade (see Figure 2.8). Almost a third of the sample (32.31%) reported that they had placed two suspects in a parade together, but more than half of the sample (58.47%) reported that they had placed three or more suspects in a parade together. These responses ranged from 3 to 36 suspects in a single parade. On average, the most suspects placed in a parade together was 4.23 (*Mdn* = 3, *SE* = 0.56).

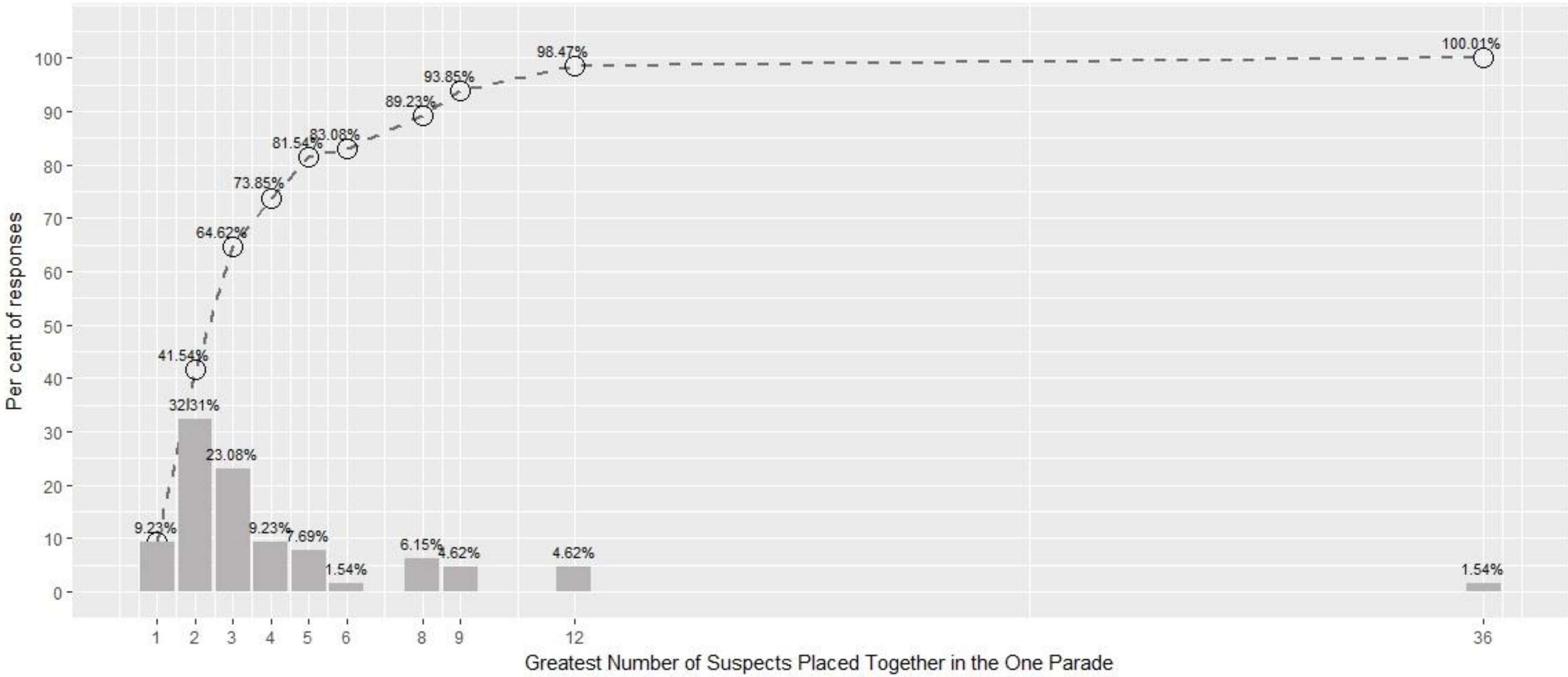

Figure 2.8 A distribution of the responses for the greatest number of suspects that officers have placed together in a single parade. The percentage values above each bar represent the proportion of the sample who reported that responses. The dashed line represents cumulative percentage, with these exact values shown next to each circular marker.

Officers were asked to estimate the largest lineup size (including foils and suspects) of any parades that they had built or administered. Inspection of the data showed that nine officers had provided the same response for (a) largest lineup size (including foils and suspects) and (b) largest number of suspects. It was not clear whether this was an error. Thus, two analyses were run: One that included this data – in case the officers had, indeed, conducted an identification parade with these numbers (some of which were as low as one) – and another without this data. For the first analysis, which included all the data, the largest parade contained, on average, 15 members (*Mdn* = 14, *SE* = 1.14). For the second analysis, where the repeated responses were removed, the largest parade contained 16.45 members average (*Mdn* = 15, *SE* = 1.14).

Logistical aspects. Police officers were asked where they found the individuals who appeared in the identification parade (i.e., the foils or bystanders) alongside the suspect (see Table 2.10). The most frequently reported response was that the foils were other detainees in the prisons or jails (40.48%). Other responses were that the suspect, instead of the police, was responsible for arranging the foils by asking friends or family members (26.19%); or that members of the public were paid or volunteered to stand in the parade (25%). Few police officers indicated that fellow police officers would stand in the parade as foils (5.92%).

Police officers were asked what type of input suspects had in the formation of the parade prior to the viewing. The most frequently reported input was that suspects could choose who would stand alongside them in the parade (41.18%). Other responses included that suspects could choose their standing position (14.71%), their clothing (7.35%), or number (2.94%). The suspect's lawyer was also allowed to provide input about the parade (16.18%).

Officers were asked to estimate the average delay between the crime and the identification parade. The most frequently reported delay was a month (21.33%), although

17.33% of officers reported that the parade would take place within a week. Some officers were unable to provide an estimate, and some felt that a parade should happen at any stage of the investigation, even if the suspect was found years after the crime occurred.

Officers reported that parades were more frequently constructed for two or more eyewitnesses (56.34%), than for single witnesses (40.85%). Two officers indicated that they constructed parades for single and multiple eyewitnesses equally (2.82%).

Of the type of parades used, most of the respondents had used a photograph parade (91.67%) and a live parade (83.33%). A smaller percentage reported that they had used a video parade (16.67%) – however, when probed, some officers (a) asked what constituted a video parade, and (b) had misunderstood a videotape parade to refer to any parade format that was recorded by a videographer from the local criminal record centre (LCRC). Officers confirmed that video parades, like the VIPER system used in the United Kingdom, are not currently used. Two officers (2.78%) reported that they had used other parades, and specified that these were voice parades.

Table 2.10.

Logistical Aspects of Administering Parades

Questions and responses	n	%	95% CI
How are foils/bystanders found?			
Detainees	34	40.48%	[29.98, 50.98]
Suspect arranges	22	26.19%	[16.79, 35.59]
Police	5	5.92%	[0.87, 10.97]
Volunteers/paid helpers	21	25.0%	[15.74, 34.26]
Legal representatives/lawyer arranges	2	2.38%	[0.00, 5.64]
Does the suspect have any input in finding/sourcing the foils/bystanders?			
Provide people for the parade	28	41.18%	[29.48, 52.88]
Standing position	10	14.71%	[6.29, 23.13]
Clothing	5	7.35%	[1.15, 13.55]
Choosing number	2	2.94%	[0.00, 6.96]
Other requests	12	17.65%	[8.59, 26.71]
Lawyer has input	11	16.18%	[7.43, 24.93]
Estimated average delay between crime and parade			
Up to one day	3	4.00%	[0.00, 8.43]
Within a week	13	17.33%	[8.76, 25.90]
Within two weeks	1	1.33%	[0.00, 3.92]
Within a month	16	21.33%	[12.06, 30.60]
Within three months	2	2.67%	[0.00, 6.32]
Within six months	3	4.00%	[0.00, 8.43]
More than six months	1	1.33%	[0.00, 3.92]
*Anytime	3	4.00%	[0.00, 8.43]
*Impossible to estimate	14	18.67%	[9.85, 27.49]
*Sooner is better	19	25.33%	[15.49, 35.17]
Are parades most frequently held for:			
Single eyewitnesses	29	40.85%	[29.42, 52.28]
Multiple eyewitnesses	40	56.34%	[44.80, 67.88]
Both	2	2.82%	[0.00, 6.67]
Used Live Parade			
Yes	60	83.33%	[74.72, 91.94]
No	12	16.67%	[8.06, 25.28]
Used Photograph Parade			
Yes	66	91.67%	[85.29, 98.05]
No	6	8.33%	[1.95, 14.71]
Used Video Parade			
Yes	12	16.67%	[8.40, 24.94]
No	60	83.33%	[75.06, 91.60]

Note. Responses denoted by asterisks were not included among closed responses for the question. Asterisked responses were given spontaneously.

Officers were asked which type of parade they preferred to use (Figure 2.9), and to explain their preference. Slightly more police officers preferred photo parades (48.28%) to live parades (42.53%). A smaller percentage preferred video parades (6.9%), but as explained previously, it appeared that officers mistook video parades as parades that were video recorded by a videographer from the LCRC.

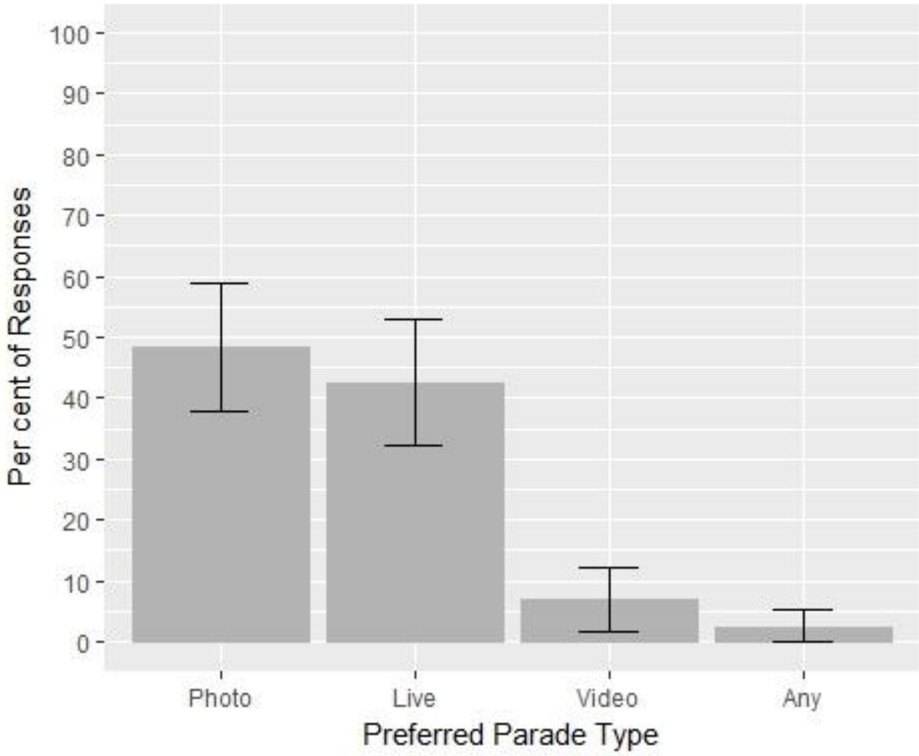

Figure 2.9 A distribution of parade formats preferred by police officers. Error bars are 95% CI.

Officers were asked to explain their parade preference responses in Figure 2.9 (Table 2.11). Of those who preferred photo parades, the most frequently provided reason was that these types of parades were easier and quicker to arrange (53.66%), followed by the belief that photo parades benefitted the witness (39.02%). Officers believed that photo parades caused the witness less distress, and that witnesses were more comfortable viewing this type of parade.

Of the officers who preferred live parades, the most frequently reported reason was that these parades helped witnesses make a ‘better’ identification (51.35%). Inspection of these responses showed that detectives thought this type of parade provided witnesses the opportunity to make requests (e.g., to ask the suspect must roll up his/her sleeves), and allowed the witness the opportunity to evaluate other information of the suspect (e.g., the height, weight, body, posture, and voice of the suspect). Almost 20% of the detectives reported that live parades were preferred by the court, and 16.22% reported that it was the more ideal parade (sometimes this was qualified by stating that this is the type of parade that they were trained in). Interestingly, 13.51% of officers responded that live parades were fairer to the suspect, since the suspect could comment on the procedure, change positions, and clothing.

Even though the SAPS does not use video parades, a few officers (n = 5) said that they would prefer video parades. Their reasons were that these videos could be used in court to demonstrate that the officer administered the parade correctly and fairly, and that videos provided more information than photo parades that facilitate recognition. However, unlike live parades -which officers also said provided more information - one officer also believed that a video parade could put the witness as ease.

Table 2.11

Motivations and Responses for Preferred Parade Format

Parade Format	**Reasons for Preferred Format**	**n**	**%**	**95% CI**
Photo	Easier to arrange, fewer logistical difficulties	22	53.66	[38.40, 68.92]
	Benefits victim	16	39.02	[24.09, 53.95]
	Familiar with method	1	2.44	[0.00, 7.16]
	Accepted by court	2	4.88	[0.00, 11.47]
Live	Aids witness to make identification	19	51.35	[35.25, 67.45]
	For legal purposes/courts	7	18.92	[6.30, 31.54]
	Most ideal	6	16.23	[4.35, 28.11]
	Fairness to suspect	5	13.51	[2.50, 24.52]
Video	For court	1	20.00	[0.00, 55.06]
	Better identification	2	40.00	[0.00, 82.94]
	Witnesses are less nervous	1	20.00	[0.00, 55.06]
	Other	1	20.00	[0.00, 55.06]

Note. Percentages sum to 100% within each preferred parade format response.

Supporting information at identification. Officers were asked five questions that probed whether eyewitnesses had to provide ancillary information about their lineup identification. These five questions, and the results, are listed in Table 2.12.

Most officers answered that eyewitnesses were required to justify their identification (82.87%). Furthermore, most officers answered that eyewitnesses were required to provide additional information about the identified person (85.92%). When asked whether eyewitnesses could name the roles of all the identified persons, the majority of officers responded that eyewitnesses were able to do this (84.81%). An example of the type of additional information given is that "He [the identified individual] was there at the door". One officer responded 'no' to this question, and when asked to substantiate their response, they said that sometimes eyewitnesses would not recall the role performed by the identified individual, but would recall who the individual had accompanied instead.

In contrast to the questions where most officers responded that eyewitnesses could name the roles and actions of all the perpetrators, 77.47% of officers also reported that eyewitnesses experienced confusion about the roles and the actions of the identified individuals. One officer reported that this happens often.

The fifth question asked whether eyewitnesses ever provided information at identification that differed from information in the first statement given to the police at the beginning of the investigation. The aim of this question was to provide an 'objective' measurement of the type of information provided by the eyewitness when making an identification: Police officers will not know whether eyewitnesses are recalling correct or incorrect information, but the statement provided by the eyewitness can act as a comparative reference. The information provided at identification could be compared to the details within the eyewitness' statement, and this would indicate whether eyewitnesses recalled the same information, new information, or contradictory information. However, compared to previous questions, the responses to this question were less clear with 50.72% and 47.83% responding 'yes' and 'no' respectively.

Officers were asked to describe how the information provided by eyewitnesses differed at identification from their statement. Their responses were grouped into four categories and coded by two coders. There was moderate agreement between the two coders, $\kappa = 0.69$, 95% CI [0.49, 0.89], $p < .001$. The most frequently reported difference was the physical appearance of the suspect (43.33%). Officers reported that the physical description provided by the eyewitnesses would differ from the physical appearance of the suspect (e.g., “in statement: suspect is dark and tall; at parade: suspect is light in complexion and short”), or that different clothing was worn at the time of the crime and at the identification (e.g., “sometimes the suspects changed clothing, therefore it's difficult for some witnesses to make positive point-outs, e.g. beanies, hoodies, etc.”). Officers also stated that sometimes eyewitnesses would remark on the change in the physical appearance of the suspect.

Equal numbers of officers reported that eyewitnesses confused the roles of the perpetrators (20%), or that they forget what they had provided in their statements (20%). Slightly fewer detectives (16.67%) reported that eyewitnesses sometimes recalled new information at the time of the parade, which was not originally included in their statement.

Table 2.12.
Responses to Whether Eyewitnesses are able to Provide Additional Information Following a Positive Identification

Questions and Responses	n	%	95% CI
Q1. Are eyewitnesses required to state a reason for why/how they made their identification			
Yes	58	82.87	[74.04, 91.70]
No	12	17.14	[8.31, 25.97]
Q2. Are eyewitnesses required to state any additional information about the person (e.g. "I recognise him because he had the gun").			
Yes	61	85.92	[78.44, 93.40]
No	10	14.08	[6.60, 21.56]
Q3. Are eyewitnesses able to name and describe the roles/actions (e.g. 'He had the gun") of *all* the people whom they identify from the parade?			
Yes	60	84.81	[76.46, 93.16]
No	11	15.49	[7.07, 23.91]
Q4. Do eyewitnesses ever seem confused about the roles and actions of the people whom they identify from the identification parade?			
Yes	55	77.47	[67.75, 87.19]
No	16	22.53	[12.81, 32.25]
Q5. Has it ever happened that the information that the eyewitness provided when making an identification differed from the information that they provided in their statement?			
Yes	35	50.72	[38.84, 62.60]
No	33	47.83	[35.96, 59.70]
Q6. How does the information provided at identification differ from that in the statement?			
Mix up suspect roles	6	20.00	[5.69, 34.31]
Forget their statement	6	20.00	[5.69, 34.31]
Description changes or mentions that the suspect looks different	13	43.33	[25.60, 61.06]
New information	5	16.67	[3.33, 30.01]

Note. The percentages for question five (Q5) do not summate to 100%, become one office responded 'sometimes'.

Discussion

The aims of the survey were to determine how South African police officers administer parades for multiple-suspect scenarios, and whether eyewitnesses are required to substantiate lineup identifications.

The results from the current survey demonstrate that police officers in South Africa use a variety of methods to build identification parades for multiple perpetrators. Almost three-quarters of the sample reported that they had at some stage administered a parade that contained more than one suspect. Furthermore, the number of suspects included in a parade was not limited to only two, as recommended by the guidelines and case law – more than half of the officers reported that they had included more than three suspects in a single parade.

Officers who preferred a single parade with multiple suspects reported that a single parade alleviated logistical difficulties and was less time-consuming to administer. Officers also believed that a single parade caused the eyewitness less emotional distress. A surprising result was that almost three-quarters of the officers responded that if the crime was committed by two perpetrators then both suspects would be included in the same parade.

It is clear from the current survey that eyewitnesses are required to provide additional information about their identification. More than 80% of police officers confirmed that eyewitnesses had to justify their identification. However, further questioning revealed that 85.92% of the officers believed that eyewitnesses could recall the roles of all the suspects whom they identified and 84.81% of police officers said that eyewitnesses could recall the roles of all identified suspects. This contrasts with 77.47% of officers who also said that eyewitnesses confused the roles performed by the identified suspects. Thus, it is not clear whether eyewitnesses can correctly pair roles with perpetrators. Unlike a laboratory experiment where the researchers can determine accuracy, police officers do not necessarily know whether the eyewitness is correct. Furthermore, determining whether the eyewitness correctly paired roles with perpetrators may not be immediately obvious or easy. The eyewitness may appear reliable by providing a detailed statement of the crime, and by identifying the suspect, however makes a mistake by assigning the incorrect actions. Determining the accuracy of role assignment is difficult, even if the actions and perpetrators are described within the statement.

Half of the respondents (50.72%) reported that witnesses provided supporting information that differed from information in their statement. Officers gave varying reasons for why eyewitnesses provided differing information. First, the eyewitness was confused and provided incorrect information (in the statement, or at the parade); second, the eyewitness recalled new information; and third, the eyewitness recalled different information (e.g., a physical description) that matches the person whom they identified rather than the person whom they saw. One officer suggested that eyewitnesses might be afraid and lie to avoid testifying (the example given was that the eyewitness provides the name of the suspect to the police when giving a statement, but then denies knowing the suspect and does not identify the suspect at the identification parade). It remains unknown whether eyewitnesses do lie or whether this is the officer's impression of eyewitness behaviour.

The police forces surveyed by Hobson et al. (2012) utilised parades delivered using VIPER or PROMAT, whereas the police stations in South Africa primarily use live identification parades. Despite the different practices utilised by the South African police services and the police forces surveyed in Hobson et al., the results from both surveys showed that police forces adapted their procedures in response to the difficulties associated with administering multiple-suspect parades. The results from the current survey of South African officers suggest that these adaptations concerned the construction of the parade (e.g., the number of suspects and foils placed in one parade), whereas the results from the U.K. survey suggested that the adaptations were mainly centred on the administration of the parade, and the instructions given to the eyewitnesses. In the current survey, I did not probe police officers about what instructions were given, and this information was not offered spontaneously by any of the officers. Hobson et al. (2012) did not report that the parade construction was adapted, that is, that more than one suspect was included in a single parade.

The survey yielded other interesting results that were not directly related to the aims. First, just over half of the officers reported that they had not testified in court, and three-quarters of the officers had testified fewer than five times. One of the detectives whom I interviewed speculated that many officers did not want to administer identification parades since they were afraid of testifying in court. Consequently, the same officers normally administer identification parades, and will do so for officers employed at different stations.

Second, over half of the current sample reported that, of the crimes that they specialised in, more than 50% were committed by multiple perpetrators. This is a higher estimate than the estimate of multiple-perpetrator crimes reported in South African statistics and research. For example, in South Africa, 50% of a small subset of participants in the Victims of Crimes Survey (Statistics South Africa, 2014, 2017) experienced a crime committed by multiple perpetrators, and between 17% and 50% of reported rapes were committed by multiple perpetrators, (Artz & Kuniski, 2003, as cited in Horvath & Kelly, 2009; Jewkes et al., 2012; Maw, 2012; Swart et al., 2000). Most international statistics hover around 20%. (Australian Bureau of Statistics, 2004; Curran & Millie, 2003; Franklin, 2004), except for crimes against minority groups, which is estimated at between 46% and 70 (European Union Agency for Fundamental Rights, 2012). The higher prevalence estimate may be due to the sample included in the current survey. This sample comprised detectives – a specialised role within the police force – who investigate more serious crimes, and more serious crimes may be more frequently committed by multiple perpetrators (e.g., rape, hijackings, cash-in-transit heists, and business robberies).

The third interesting finding was the reasons provided for the different parade procedures. Officers reported that both types of parades (single-suspect, multiple-suspect) were beneficial for the witness, albeit for different reasons. Multiple-suspect parades were considered less distressing for eyewitnesses, since eyewitnesses had to view only one parade. In contrast, single-suspect parades were considered easier for eyewitnesses, because eyewitnesses had to make only one decision per parade. Furthermore, the single-suspect parade was considered fairer to the suspect. Officers mentioned some concern over whether identification parades would be accepted in court, and whether the parades would be easy to defend. These comments may support the hypothesis that officers are afraid to testify and for this reason avoid administering identification parades.

Fourth, the finding that foils are primarily detainees or jail inmates supports findings from Wogalter, Malpass, & Mcquiston (2004). In their survey, police officers reported that foils for live parades were primarily sourced from the jails (79%), or were police officers (60%) or volunteers (37%). For the current survey, the South African detectives reported that it was difficult to find volunteers to stand as foils in the identification parades, because volunteers were required to be available for long periods of time, and were too afraid to participate. Approximately a quarter of the current sample (26.19%) responded that the suspect was responsible for arranging the foils. When pressed for an explanation, some officers suggested that it was more difficult for the defence to argue in court that the lineup was unfair when the onus of ensuring a fair parade shifted from the police to the suspect and their lawyer.

When asked about parade preference, approximately 48.28% of officers preferred photo parades, whereas 42.53% preferred live parades. One reason given for the preference for photo parades was that eyewitnesses and victims would express less distress if the suspect did not appear in-person. This finding contrasts with results from Hobson et al. (2012) where police officers reported that eyewitnesses were distressed even when they viewed video parades. Unfortunately, police officers may not be able to avoid upsetting witnesses by using different identification parade formats. It remains unanswered whether the emotional distress experienced by eyewitnesses and victims is less for some parades (e.g., photo or video) than for others (e.g., live parades).

Officers who preferred live parades felt that these parades provided more opportunities for the eyewitness to make an identification. For example, the eyewitnesses could request that the suspect walks or talks. One of the participating officers provided anecdotal evidence about an eyewitness who, while viewing the parade, suddenly remembered that the perpetrator had a tattoo inside their bottom lip. The administrating officer requested that each person in the parade show the inside of their lower lip. This request could not have been fulfilled with a photo parade, but could be fulfilled at a live parade. Extent psychological literature suggests that dynamic parades (such as live parades, and video parades) facilitate better recognition than photo parades (live parades: Cutler, Fisher, & Chicvara, 1989; Egan et al., 1977; Kerstholt, Koster, & van Amelsvoort, 2004; video parades: Havard, Memon, Clifford, & Gabbert, 2010; Valentine, Darling, & Memon, 2007), but some research demonstrates that video parades result in better recognition performance than live parades (Valentine & Heaton, 1999; Valentine, Harris, Piera, & Darling, 2003). Live parades can provide the eyewitness with opportunities to view additional information – as in the anecdotal example - but it is unclear how frequently eyewitnesses request such opportunities. Further research is needed to establish (a) whether recognition performance is better following live, video, and photo parades, and (b) whether eyewitnesses are less distressed when viewing live, video, and photo parades.

One pivotal difference between photo parades and live parades is the amount of input that the suspect and their lawyer have on the construction and administration of these parades. Of the officers who preferred live parades, 13.51% reported that live parades were fairer to the suspect, because the suspect was legally allowed to give input to the construction of the parade. In contrast, suspects are not entitled to provide input to photo parades. For example, the guidelines identified in Du Toit's commentary state that the suspect's lawyer must be present during a formal parade (i.e., a live parade; see Rule 3, 11, and 12 in Appendix B), but the lawyer need not be present for the photo parade. The absence of legal counsel during the administration of the photograph parade is supported by case law: The suspect is not entitled to legal representation during a photograph parade. [43] The lack of guidelines surrounding photo parades prompts further investigation for future research.

[43] S v Hlalikaya and Others 1997 (1) SACR 612 (SE)

Limitations

As with most self-report measures, there is a possibility that the responses given to the current survey were not a true reflection of how police officers administer identification parades in practice. Furthermore, to avoid criticism, police officers may have given answers that were legally or procedurally correct (as per case law and guidelines). This was evident when I probed officers about some of their responses. For example, when I asked officers what they considered a reasonable number of foils that they could arrange for a parade, most officers gave the recommended number of foils as stated in the guidelines. The recommended number of foils, however, may not be easy to find. Additionally, for some questions officers were asked to provide estimates for their answers, and these estimates may be erroneous (e.g., the average delay between the crime and the parade). I am not implying that police officers are purposefully providing incorrect responses; however, I think that it is extremely important to establish rapport and trust with detectives and detective commanders, and encourage a research relationship with them before and after administering the survey. To this end, I intend to disseminate the results of this survey in the form of a report to the participating police stations, and to the head of research in SAPS, so that they are aware of the findings. While this may not benefit me directly (especially since the current survey is now complete), I hope that this continues to foster a relationship between SAPS and UCT.

A second limitation is that I did not report any supporting information that confirmed or disconfirmed the police officers' responses. Future research could consider (a) obtaining photographs of identification parades, and (b) case dockets of completed cases; there is a risk that this information might be biased. I tried to request this information from SAPS, but they were unwilling to provide any information about a case where the suspect was found innocent or the suspect was not yet tried. Thus, the remaining dockets and lineup photographs would be of cases where the suspect was guilty, and these data could be biased. Instead, future researchers could consider attending parades and capturing the parade details. This could yield insight into how actual parades are formed. A final suggestion would be to interview or survey lawyers who work for the National Prosecuting Authority (NPA). These lawyers could provide information about successful and unsuccessful cases, and they may have experience and feedback about how lineup administration in practice affects the course of an investigation and a case at trial.

A third limitation is that the majority of my sample had not constructed an identification parade within the last year, and 42.67% of the sample had constructed five or fewer parades in their entire career. Consequently, the conclusions that can be drawn from this sample are limited as most of the sample was inexperienced with identification parades, and it is not known why they had conducted so few parades in the last year or in their career.

Conclusion

Despite the limitations, the findings of the current survey give insight into actual police practice and contribute to the existing body of knowledge about how police officers administer identification parades. Furthermore, the finding also highlights the practical difficulties that police officers experience when conducting identification parades, and police officers' preference for certain types of parades (and their reasoning thereof).

Additionally, the results show that most police officers have at some stage of their career administered a parade that included more than one suspect, and that approximately a third of officers preferred to administer multiple-suspect parades. Officers also reported that eyewitnesses are expected to supplement their identifications with additional information, but that eyewitnesses get confused and are not able to do this. This finding, however, is difficult to interpret: Officers do not know if eyewitnesses are correct about either the identification or the pairing, and they may not even have considered whether the pairing was correct. Case law, however, suggests that the courts believe that eyewitnesses can make this pairing,[44] and the ability to correctly pair perpetrators and roles attests to the veracity of the eyewitness' memory. What remains unanswered is whether eyewitnesses can accurately pair perpetrators with roles, and whether the role-perpetrator pairing (i.e., associative memory) is affected by the number of perpetrators observed at encoding. In the next chapter, I will review the relevant literature for associative memory for faces and information, and discuss how set size at encoding impacts memory.

[44] Supra footnote 4

Chapter 3

Recognition Memory for Multiple Faces

In the previous chapter, I established that police officers ask eyewitnesses to substantiate their identification decisions with supporting information. Some officers reported that eyewitnesses were not always able to name the roles performed by each perpetrator, or that they confused perpetrator roles with one another. What remains unanswered, however, is (a) the number of perpetrators eyewitnesses can recognise, and (b) whether eyewitnesses can correctly pair associated information (such as the role) with each perpetrator?

In the current chapter, I review the face recognition literature to determine (a) how many faces humans can recognise, (b) how set size affects face recognition, and (c) how associative memory (e.g., between faces and semantic information) is affected by set size at encoding. I focus on the relevant face recognition literature where set size was manipulated at encoding as a precursor to Chapter 5, where I will review the relevant eyewitness literature that investigates eyewitness memory for multiple perpetrators. Research on face recognition and eyewitness memory are complementary, answering basic and applied questions on the same topic. With different methods and materials these two research areas answer the same question, that is, how well and under what conditions we can remember human faces. Despite the differences between the two research areas, they inform each other: Face recognition research is the theoretical precursor to eyewitness memory experiments, which is, in turn, its applied extension.

Literature Review: Human Ability to Recognise Faces

To begin the investigation into what can be expected of eyewitness memory for human faces, I will first review the literature examining human face recognition. I will then examine the literature comparing faces to 'general' visual images, followed by a review of the literature on recognition of familiar and unfamiliar faces. Following this, I will review research in which faces were paired with other information at encoding (e.g., names).

Face Recognition versus Eyewitness Memory Studies

Face recognition experiments share some similarities with eyewitness memory experiments: Both normally involve a target face, which must be studied and recognised at a later stage. There are, however, a few notable differences between these two research areas (Shapiro & Penrod, 1986). Face recognition experiments normally incorporate multiple targets whereas eyewitness memory experiments use only one target. This is most likely a consequence of face recognition experiments evolving from research in psychophysics where multiple stimuli are presented to a participant so that stimuli bias (e.g., distinctiveness) is neutralised (Howells, 1938). Thus, the use of multiple targets within face recognition experiments is a methodological artefact rather than a strategic choice to answer a research question about set size. Additionally, in face recognition research, faces are often presented without context whereas in eyewitness research the target 'face' is of the perpetrator of a crime. Target faces are normally, but not always, presented as images in face recognition research, whereas there is more variation in stimuli presentation within the eyewitness literature; for example, target faces might be encoded as images (e.g., Loftus, 1976), videos (e.g., Brewer & Wells, 2006), or in live encounters (e.g., Ihlebæk, Løve, Eilertsen, & Magnussen, 2003). The encoding context in eyewitness research is more visually complex than in face recognition research and has an accompanying narrative about the crime. A third difference between these two research areas is the type of recognition test used. In face recognition research, memory is normally tested with an Old-New task or an Alternative-Forced Choice task (AFC; often referred to as a two-alternative forced-choice task), whereas in eyewitness research, memory is typically tested with an array of images, that is, a lineup task. In Old-New tasks, all or a subset of the previously studied (i.e., Old) faces are presented sequentially, with previously not-studied (i.e., New) faces interspersed between them. Participants must decide for each face whether it was seen previously or not. In an Alternative-Forced Choice task, normally two faces (one Old, one New) are shown simultaneously, and participants must decide which of the two faces was studied previously. In contrast, an image array, like that used in eyewitness research, is meant to simulate the recognition task experienced by eyewitnesses when viewing an identification parade. The array (or lineup) comprises of multiple individuals (more than two, and closer to six or eight) who physically resemble one another, and are chosen based on either (a) their visual resemblance to the target face in the parade, or (b) their resemblance to the verbal description of the target face in the parade (for a discussion on these two selection methods, see Clark & Tunnicliff, 2001; Luus & Wells, 1991).

The differences and similarities between face recognition and eyewitness memory experiments are important to consider, because they yield different results and answer different research questions. In a meta-analysis on face recognition, Shapiro and Penrod (1986) included both eyewitness and face recognition experiments, and acknowledged that the methodological differences between these two areas obfuscated the results; see their meta-analysis for an overview of the similarities and differences between face recognition and eyewitness memory experiments). It is for this reason that I am highlighting the differences between the two research areas, while noting their mutual contribution to what is known about memory for faces. From this point onwards in this chapter, the review focuses on face recognition research.

Limits of Human Memory for Visual Items

The first question that this review addresses concerns the limits of human memory for faces: How many faces are humans able to remember? To answer this, I briefly review the literature on the upper limits of memory for non-verbal and verbal stimuli, of which faces could be a considered a special type of visual item (Yue, Tjan, & Biederman, 2006).

Memory for visual stimuli is superior to memory for sentences, and words, as demonstrated by Shepard (1967) who compared recognition performance on these three types of stimuli. Participants studied either 540 words, 612 sentences, or 612 pictures; after studying the respective stimuli, participants completed a two-alternative forced-choice (2AFC) task. Despite the large set size at encoding, participants responded to a smaller subset of the entire learning set instead (60 2AFC pairs for words, and 68 2AFC pairs for sentences and pictures respectively). The results showed near-perfect recognition abilities: Participants were most capable of recognising previously studied pictures (*Mean* = 96.7%), but also displayed excellent memory retention for sentences (*Mean* = 89.0%) and words (*Mean* = 88.4%). Shepard (1967) also manipulated the delay between encoding and recognition of pictures (either 2 hours, 3 days, 7 days or 120 days), and found that accuracy remained above 85% for up to seven days after encoding. Only at 120 days after encoding did recognition performance drop to 57.7% accuracy for pictures.

Since then, studies have notably increased the set size of visual items at encoding to 200 images (Nickerson, 1965), 1 000 and 2 560 items (Standing, Conezio, & Haber, 1970), 2 500 images (Brady, Konkle, Alvarez, & Oliva, 2008), almost 3 000 images (Konkle, Brady, Alvarez, & Oliva, 2010b) and 5 000 images (Konkle, Brady, Alvarez, & Oliva, 2010a), and up to 10 000 items (Standing, 1973). The overwhelming consensus from these studies is that memory capacity (both immediate and long-term) for visual items is large and seemingly unbounded. Accuracy scores are often upwards of 80%, for example, Nickerson (1965) reports average accuracy results of 95%. Retention of visual images appears to withstand the deleterious effect of time, up to a week (Vogt & Magnussen, 2007) or almost a year later (Nickerson, 1968). The immunity to delay is stronger for visual items that were studied more than once, even if that encoding was incidental (Nickerson, 1965). For example, Nickerson (1965) tested 56 participants on their recognition memory for 200 visual images, each image shown one at a time for 5 seconds at encoding, with a 2AFC task comprising 200 pairs of images. On average, participants achieved 95% accuracy. Nickerson (1965) hypothesised that accuracy would decrease as the delay and/or interference between encoding and test increased – results showed that performance did drop as the number of interfering images increased, but this decrease was minimal, from approximately 97% when 40 items were presented between encoding and test to 87% when 200 items were presented between encoding and test.[45] Furthermore, Nickerson (1968) introduced an extension of his original experiment reported in 1965: The participants who were first tested in Nickerson (1965) were retested on 200 images either 1, 7, 28 or roughly 360 days later. The 200 images used at recognition comprised 100 New items, and two sets of 50 Old items each. The first set of Old items was first encoded during the encoding stage in Nickerson (1965), where it also appeared as an Old item during recognition; therefore, participants had had two encoding opportunities for this material. The second set of Old items was not encoded during the encoding stage, but was presented as New items during the recognition phase of Nickerson (1965); consequently, participants had one incidental encoding opportunity for these stimuli. Overall, Nickerson (1968) showed that recognition performance was consistently superior by roughly 20% for double-encoded Old items than single-encoded Old items. Across the four delays, accuracy for double-encoded Old items decreased from approximately 98% to 62%, and accuracy for single-encoded Old items decreased from approximately 78% to 37%.[46] Surprisingly, the false alarm rate remained consistently low: it was less than 10% up to 28 days after encoding, and then roughly doubled to about 20% at 360 days post encoding.

[45] Nickerson (1965) does not report exact percentage values for these data, so these approximate values were read from Figure 1.

[46] Nickerson (1967) does not report exact percentages here and these values were estimated from Figure 1.

Overall, these results suggest that recognition memory for visual items is impressive and if face images are considered a form of visual stimuli, then these results suggest that capacity limits for faces are vast. There are some methodological differences between the reviewed research and research on face recognition that are worth noting and that may impact extrapolation of these findings. The literature reviewed up to this point tested memory for visual items using a 2AFC test. AFC tasks are different to Old-New tasks typically used in face recognition research: In an AFC task, participants decide which of the two options was previously encoded, knowing that one of the answers is correct. In comparison, in an Old-New task, participants must make a decision for each stimulus, not knowing whether that item is old or new. Thus, for Old-New tasks, participants' responses demonstrate both memory performance and response criterion. Response criterion captures choosing behaviour and the willingness of participants to commit to a decision (Macmillan & Creelman, 2005); participants with high criterion are less willing to recognise an item as Old (i.e., there will be more New responses) whereas participants with low criterion are more willing to state that an item is Old (i.e., there will be fewer New responses). While it is possible to calculate response bias statistically from 2AFC tasks (see DeCarlo, 2012), theoretically the concept of response bias does not exist because – as stated in the name of the task – participants are forced to choose between two items irrespective of their criterion. For this reason, 2AFC tasks do not capture all the characteristics of signal detection measures, and outcomes from a 2AFC task may be inflated when compared to an Old-New task.

Furthermore, the estimates of memory capacity from the reviewed research may be inflated since recognition was only tested for a subset of the encoding sample (Brady et al., 2008; Konkle et al., 2010a; Konkle et al., 2010b; Nickerson, 1968; Standing et al., 1970; Standing, 1973; Vogt & Magnussen, 2007). Only Nickerson (1965) tested recognition memory for the entire encoding set. The properties of the sample of images used at testing are unknown

– was this sample difficult, easy, or representative of the entire set? For the reviewed experiments, the testing sample was chosen randomly from the encoding sample, but Wittwer (2018) highlights that characteristics of encoding and test stimuli are extremely important and there is currently no standardized method to ensure that stimuli are equally difficult.

While these studies provide evidence of the upper limits of human memory for visual items, human faces – while visual – are a different type of visual item. The discussion surrounding the properties that make human faces unique and different from other visual items is beyond the scope of this thesis (but see Barry, Johnston, & Scanlan, 1998; Diamond & Carey, 1986; Farah, 1996; Farah, Wilson, Drain, & Tanaka, 1998; Kanwisher, 2000; Kanwisher, McDermott, & Chun, 1997; Liu & Chaudhuri, 2003; Moscovitch, Winocur, & Behrmann, 1997; Tarr & Gauthier, 2000; Yue et al., 2006), although there are some qualities that should be briefly mentioned, especially pertaining to methods used to test visual memory. The aim of face recognition and eyewitness memory research is to test memory for the face and not for the picture, as first stated by Bruce (1982). Bruce argues that in an applied setting, face recognition would be demonstrated by recognising an individual in another context, in a different outfit, from a different view/pose/position, and even from a different point in time from the original encoding. Face recognition is not demonstrated when the same image of an individual is used at both encoding and recognition, because participants may make a correct identification due to an artefact of the photograph. To control for this and introduce variation, face recognition researchers use different images of the target at encoding and recognition, or digitally edit the image (e.g., transform the image into greyscale). For face recognition experiments, it is possible to introduce variation between encoding and recognition, but is it possible (or meaningful) to do so in an experiment testing memory for pictures? One solution would be to have two photographs from the same scene, but from different orientations; one of these photographs could be used at encoding, and another at recognition. If, however, researchers use two different pictures, are they still measuring memory for visual items, or are they measuring item context? Furthermore, Vogt and Magnussen (2007) demonstrated that the peripheral and seemingly unimportant details for an image, for example, lamp posts on the side of a door (see Vogt & Magnussen, 2007, p. 6), were necessary for superior picture recognition and added to the better encoding of these stimuli. The final difference between faces and static visual images is that there is natural within-face variation. Consider a photograph of a stapler and a photograph of your face. The 'state' of your face is dynamic: You can pull different expressions, your face moves when you speak, it changes colour depending on whether you have exercised or if you are cold, and it changes with weight gain and age. There is no 'natural' 'resting' state of your face. In contrast, the state of a stapler is not likely to change. It is a consequence of this natural variation within the same faces which means that recognition of faces is more elusive (see the discussion about structural and pictorial codes in Bruce & Young, 1986). The difference in the research methods of these two areas is a consequence of the theoretical argument underpinning both areas: Face recognition is more complicated than image recognition, because face recognition researchers aim to measure identity recognition.

Memory for Familiar Faces

It is well known from anecdotal evidence that human memory is almost limitless for familiar faces. We have anecdotal evidence of this from our personal lives, because we can recognise family members, loved ones, friends, and colleagues across a variety of contexts, time periods, and outfits. Recognition of familiar faces is especially robust to changes in time, perspective, outfit, and aging (for a review see Johnston & Edmonds, 2009). Some studies that have investigated facial recognition for familiar faces (Bahrick, Bahrick, & Wittlinger, 1975), and famous faces decades after they were first encountered (Bäckman, 1991; Bäckman & Herlitz, 1990), found that older adults were capable of recognising familiar and famous faces with at least 70% accuracy. Familiar face recognition is immune to the deleterious effects of time: Bahrick et al. (1975) showed that participants recognised pictures and names of old classmates with 73% and 77% success, respectively, even 47 years after graduation.

Memory Limits for Unfamiliar Faces

Testing recognition for familiar faces is not completely analogous to the recognition task expected of eyewitnesses who must identify an unfamiliar person. However, from this research, we know that memory for familiar faces is resistant to the deleterious effects of delay, and unbounded (Bäckman, 1991; Bäckman & Herlitz, 1990; Bahrick et al., 1975).

Has any research systematically tested the limits of human memory for unfamiliar faces? Shapiro and Penrod (1986) published a meta-analysis on facial recognition, which included 128 manuscripts (190 studies with 960 experimental conditions), and examined the effect of various experimental conditions on facial recognition performance. The most relevant variables coded in the meta-analysis to my line of inquiry are (a) the number of targets present both at study and recognition, (b) the total number of faces at study (including target faces and distractor faces) at study, and (c) the number of foils (i.e., non-target faces) at recognition. For each of these three variables, Shapiro and Penrod reported high averages of faces (M = 22, M = 25, and M = 40 respectively); the averages were high due to including face recognition experiments among the coded studies. When analysing the study characteristics, Shapiro and Penrod found that three primary clusters of variables explained most of the variance: Optimality of Viewing, Load at Study, and Load at Recognition. Of these three clusters, Load at Study and Load at Recognition are most relevant to the present research. Load at Study describes factors that affect the cognitive load of encoding, whereas Load at Recognition describes the factors that affect the cognitive load at retrieval. The following three study characteristics contributed to the first cluster, Load at Study: number of target faces present at encoding, the total number of faces (i.e., targets and foils), and the duration of encoding time. Interestingly, both Hits and False Alarms increased as Load at Study increased. Due to the simultaneous increase in both outcome variables, Shapiro and Penrod (1986) suggest that Load at Study caused participants to adopt a more liberal response criterion. Furthermore, there was a negative relationship between *d'* and load at encoding. [47]

The second cluster, Load at Recognition, also comprised three characteristics. the number of foils present at recognition; number of total faces (i.e., target faces and foils) at recognition; and the ratio of targets to foils at recognition. Although Load at Recognition had no effect on Hits, it did have a significant effect on False Alarms. Specifically, False Alarms decreased when either (a) the number of simultaneously presented faces at recognition or (b) the number of foils increased. In sum, it appears that Load at Encoding decreases *d'* by increasing False Alarms more than Hits, and Load at Recognition (which is a consequence of higher set sizes at encoding) decreases False Alarms. [48]

In the studies discussed up to this point, the effect of set size on encoding was a consequence of experimental design, and not the subject of a research question. There are, however, three published studies that explicitly tested the effect of set size at encoding on face recognition (Lamont, Stewart-Williams, & Podd, 2005; Metzger, 2002; Podd, 1990). In the earliest experiment, Podd (1990) manipulated two variables, memory load and delay, between participants. Participants were shown either 20, 35, or 50 Photofit faces[49] at encoding (referred to as Set Size 20, 35, and 50 respectively), and delay between encoding and recognition was

---

[47] Discriminability, or *d'*, is calculated from the following formula *d'*= z(proportion of Hits) – z(proportion of False Alarms. The middle point for *d'* is zero, with higher values indicating better discriminability.

[48] The results of this meta-analysis are difficult to interpret, because it is not clear how the different types of studies (eyewitness identification versus facial recognition) interact with these study characteristics.

[49] These are not photographs of real individuals, but instead are constructed using Photofit, a face composite method that was used by the police (see Ellis, Davies, & Shepherd, 1978 for more information about PhotoFit).

either 10 minutes, one week, or two weeks. Following the delay, participants were shown an equal number of New and Old faces. A consequence of larger set sizes at encoding is that there will be larger set sizes at recognition: Participants must respond to more stimuli at recognition, and this increased load at recognition could affect recognition accuracy (as demonstrated by Shapiro & Penrod, 1986), for example, increased load at recognition might cause participants to adopt a more conservative decision threshold as the recognition test progresses.[50] To control for effect of load at recognition, Podd introduced an interesting method: Regardless of which experimental group participants were assigned to, the first twenty trials at encoding and the first forty trials at recognition contained the same faces for all three groups, and then the next fifteen trials at encoding and the next thirty trials at recognition contained the same faces for the Set Size 35 and Set Size 50 groups. The remaining 15 faces at encoding, and 30 faces at recognition were unique to the Set Size 50 group. As a result, when comparing recognition performance on the first forty trials, the only difference between the three groups was the load at encoding. This design is shown in Table 3.1.

Table 3.1

Experimental Design Utilised by Podd (1990) to Control for Load at Encoding and Recognition

		Encoding			Recognition		
Group	Load at Encoding	First 20 trials	Next 15 trials	Last 15 trials	First 40 trials	Next 30 trials	Last 30 trials
1	20	Set 1	-	-	Set 1	-	-
2	35	Set 1	Set 2	-	Set 1	Set 2	-
3	50	Set 1	Set 2	Set 3	Set 1	Set 2	Set 3

Note. The differences between experimental groups was load at encoding.

Overallresults showed a detrimental effect of load at encoding on *d’*. That is, as load at encoding increased, so recognition performance decreased. The negative effect on *d’* was

[50] However, note that this finding was interpreted to mean that large arrays of simultaneously presented faces - a method normally used in eyewitness experiments - increased decision criterion, and it remains unclear whether this same interpretation holds for large numbers of sequentially presented faces, like in face recognition experiments.

primarily due to a decrease in Hit Rate, whereas False Alarm Rate remained unchanged. These results are shown in Table 3.2.

Table 3.2

Hit Rate, False Alarm Rate, and d' for the Three Experimental Groups in Podd (1990)

		Recognition								
		All Trials			Trials 1 - 40			Remaining Trials		
Group	Load at Encoding	HR	FAR	*d'*	HR	FAR	*d'*	HR	FAR	*d'*
1	20	.74	.20	1.55	0.74	0.20	1.55	-	-	-
2	35	.68	.23	1.32	0.67	0.22	1.32	0.70	0.24	1.33
3	50	.65	.24	1.14	0.67	0.24	1.25	0.64	0.24	1.13
	Overall average	.69	.22	1.34	0.69	0.22	1.37	0.67	0.24	1.23

Note. HR and FAR denote Hit Rate and False Alarm rate respectively. The values for all trials, and trials 1-40 are taken directly from the tables in Podd (1990). The values for the remaining trials were calculated from Table 3 in Podd (1990).

To isolate the effect of load at encoding while controlling for load at recognition, Podd analysed the initial forty recognition trials between the three load-at-encoding groups. The first forty recognition trials contained the same forty faces for all three experimental groups. His results suggested that accuracy decreases with as load at encoding increased. The decreased recognition performance (as illustrated by a decreasing *d'*) was a consequence of a significant decrease in Hits, while False Alarms remained largely unchanged.

Podd then compared the difference in HR, FAR, and *d'* means between the first forty recognition trials and the remaining recognition trials. The only difference that was worth commenting on – and which is not presented here since these means were collapsed across delay – was that participants in the Set Size 50 group who experienced a 10-minute delay performed better for the first forty trials compared to their remaining trials. There were no significant differences between the recognition performance of three experimental groups for the remaining trials, although the results in Table 3.3 suggest a trend where HR and *d'* were lower for the Set Size 50 group compared to the Set Size 35 group. Podd neglected to report beta or criterion so it is not clear whether set size changed participants' response bias.[51]

Metzger (2002) investigated the learning abilities of adults and children while controlling for load at encoding and recognition. He recruited 180 participants (36 children, 36 young adults, and 36 senior adults) and randomly assigned them to one of three set size groups that studied either 10 faces, 20 faces, or 30 faces. Memory was tested with an Old-New test. Like Podd (1990), Metzger used artificial images; these images were black and white[52] photorealistic images of faces generated using ComPhotoFit.[53] Metzger employed a similar method to Podd (1990) to control for load at recognition: Images at encoding were divided into three sets of 10 (total images = 30), and images at recognition were divided into three sets of 20 (10 old and 10 new; total images = 60). All participants encoded the same 10 images for the first 10 trials at encoding, and participants in the two larger set size groups then studied the same second set of 10 faces. At recognition, the first twenty Old-New trials comprised the same twenty faces for all participants, followed by the next twenty Old-New trials that used the second set of images that were shown to the two larger set size group (see Table 3.3)

[51] I attempted to replicate the *d'* values reported in Podd (1990) with the Excel functions outlined in Todorov and Stanislaw (1999), where *d'* is calculated from *normsinv(hits) – normsinv(false alarms).* With this quick, crude analysis, I was unable to replicate the *d'* values reported in Podd (1990). A possible explanation for this that *d'* is an intrasubject measurement (i.e., *d'* is calculated at the subject-level). For this same reason, I am unable to reliably calculate bias and criterion from these data.

[52] Newer versions of ComPhotoFit now produce colour images.

[53]http://www.sirchie.com/forensics/crime-scene-documentation-photography/facial-composite-sketching.html

Table 3.3
Experimental Design Utilised by Metzger (2002) to Control for Load at Encoding and Recognition

		Encoding			Recognition		
Experimental Group	Load at Encoding	First 10 trials	Next 10 trials	Last 10 trials	First 20 trials	Next 20 trials	Last 20 trials
1	10	Set 1	-	-	Set 1	-	-
2	20	Set 1	Set 2	-	Set 1	Set 2	-
3	30	Set 1	Set 2	Set 3	Set 1	Set 2	Set 3

Metzger found a significant main effect for set size on *d'*: Participants in the lowest encoding group (Set Size 10) scored a significantly higher *d'* (*d'* = 2.05) than participants in both Set Size 20 and Set Size 30 groups (*d'* = 1.54 and *d'* = 1.23 respectively), who did not differ from each other. There was also a main effect for age, with children performing significantly worse (*d'* = 1.29) than both adult groups, but the adult groups did not differ from each other (*d'* = 1.64 versus *d'* = 1.88 for young and senior adults respectively). There was no interaction between set size and age on *d'*.

The significant difference in *d'* between the Set Size 10 group and the remaining two groups is driven by a significant increase in Hit Rate for the Set Size 10 group compared to the other two groups, and by a significant increase in False Alarm Rate for the Set Size 30 group. The Set Size 10 group achieved the highest Hit Rate (.78) compared to the Set Size 20 (.72) and Set Size 30 (.67) groups, and the latter two groups did not differ from each other. In comparison, Set Size Group 30 had the highest False Alarm Rate (.25) compared to Set Size Group 10 (.17), and the False Alarm Rate of the Set Size Group 20 (.20) did not differ significantly from either of the other two groups.

Metzger further expanded on these results by including calculations and analyses on *c*, a measure of response criterion (Macmillan & Creelman, 2005); however, there were no significant main effects nor a significant interaction on *c*. For this reason, Metzger concludes that set size at encoding does not change response bias, but instead, alters memory – this, in turn, affects the outcomes variables HR, FAR, and *d'*.

In the third experiment that manipulated set size at encoding, Lamont et al. (2005) added another factor: congruency between target age and participant age. Ninety-six participants (32 young adults, 32 older adults, and 32 much older adults) were randomly assigned to a low-memory-load (LML) group or a high-memory-load (HML) group and studied 20 faces or 40 faces. Unlike Podd (1990) and Metzger (2002), the target images were colour images of real individuals. All the photographs were frontal-view facial photographs of men depicting a neutral expression, but subjects in half of the photographs were college-aged students and in the remaining half, subjects were senior adults. Set size at recognition was controlled using the same technique as Metzger (2002) and Podd (1990): The first 20 faces at encoding and the corresponding 40 faces at recognition were the same for both the low-memory-load and high-memory-load groups.

A significant main effect of set size on *d'* was found across all trials. The LML group outperformed the HML group (*d'* = 2.07 versus *d'* = 1.75 respectively). Unlike the previous pattern of findings reported in Podd (1990) and Metzger (2002), the significant difference in *d'* was not a result of significant differences in HR between the two groups (HR was .78 for both groups); instead, *d'* was a consequence of a change in FAR. Participants in the HML group made significantly more false alarms (.22) than participants in the LML group (.19). An unexpected outcome was that the authors found no significant difference in *d'* between the HML and LML groups when they isolated the first 40 trials shown at recognition. The study characteristics and overall results of the three studies discussed here are presented in Table 3.4.[54]

[54] Lamont et al. controlled for the own-age bias (Rhodes & Anastasi, 2012) by showing participants photographs of both young adults (20-26 years old) and older adults (63 – 97 years old). There was a significant interaction between participant age group (younger than 40 years old, 60 – 75 years old, older than 75 years) and the age group of the faces in the photographs (Young,Old): The own-age bias was present for only the two older adult participant groups, and was not present for the younger adult participant group. The own-age bias may have interacted with load, although the descriptive statistics suggest that overall recognition was better for all three groups for LML, and worse for HML.

Table 3.4

Study Characteristics and Results for Podd (1990), Metzger (2002), and Lamont et al. (2005)

	Study Characteristics					Results				
Study	Encoding Material	Encoding time	Delay	Recognition Task	Recognition Characteristics	Encoding Set Size	HR	FAR	*d'*	*c*
Podd (1990)	Greyscale, photofit faces	5 seconds per image, 3 seconds ISI	10 min, 1 week, or 2 weeks	Old-New	Conditions not mentioned	20 35 50	.74 .68 .65	.20 .23 .24	1.55 1.32 1.14	- - -
Metzger (2002)	Greyscale, ComPhotoFit faces	5 seconds per image, 2 seconds ISI	Delay not mentioned	Old-New	Image shown for 5 seconds, 2 seconds ISI	10 20 30	.79 .72 .67	.17 .20 .25	2.05 1.54 1.23	.09 .16 .14
Lamont et al. (2005)	Color, photographs	5 seconds per image, 3 seconds ISI	No delay	Old-New	Image shown for 5 seconds, 3 seconds ISI	20 40	.78 .78	.19 .22	2.07 1.75	.02 .003

Note. ISI refers to interstimulus interval. HR denotes Hit Rate, and FAR denotes False Alarm Rate. Podd (1990) does not mention whether the recognition trials had time constraints. Podd did not calculate any measures for *c*. Metzger (2002) did not mention whether there was any delay between encoding and recognition, whereas Lamont et al. (2005) stated that there was no delay. All three studies isolated the effect of load at recognition as described in the text. Neither study introduced any variation in stimuli between encoding and recognition.

Inspection of Table 3.4 highlights the similarities and differences between the three studies. In all three studies, encoding was set to 5 seconds per image, and load at recognition was controlled for by isolating the first set of images seen at encoding and recognition so that they were the same for all experimental groups. Additionally, all three studies tested recognition using an Old-New task. There are, however, a few notable differences. First, only one study (Lamont et al., 2005) used real photographs as stimuli. However, Lamont and colleagues did not control for picture recognition (see Bruce, 1982), and like Podd (1990) and Metzger (2002), used the same images at encoding and recognition. Second, only Podd mentioned that there was a planned delay between encoding and recognition; this delay was manipulated between participants. Lamont et al. (2005) did not introduce a delay, and Metzger (2002) did not mention delay.

In all three studies, set size at encoding had a significant, negative impact on *d'*, with higher set size groups performing worse than lower set size groups. The underlying pattern of results with Hits and False Alarms differed between the three studies. Podd (1990) found that increased set size reduced HR, while FAR remained unchanged. In contrast, Metzger (2002) reported a main effect of set size on both Hits and False Alarms: As set size increased, HR decreased whereas FAR increased. Finally, Lamont et al. (2005) found no main effect of set size on HR – instead, the decrease in *d'* with higher set sizes was due to an increase in FAR. None of the studies found a significant main effect of set on *c*, which suggests that the change in recognition performance was a consequence of changes in memory, rather than response bias.

Testing for Associative Memory

So far, the reviewed literature has manipulated set size for faces at encoding and recognition. This partially answered two of the questions posed at the beginning of the chapter: How many faces are people able to remember, and how does set size affect recognition? What remains unanswered is how set size affects memory for paired, associative memory. Neither of the three studies reviewed in the previous section included any associated information with the faces at encoding.

Associative memory differs from item memory. In the studies reviewed thus far, participants were tested on *item* memory, that is, memory for only faces, pictures, or words. In contrast, associative memory tests the connections or relationships *between* items – that is, whether two items were studied together. These items can be (a) the features or components of a larger test item, for example, the features of a face or a composite face (Young, Hellawell, & Hay, 1987) or parts of a compound word (Reinitz & Demb, 1994); (b) items from the same domain, for example two faces (Reinitz & Hannigan, 2001); or (c) items from different domains, such as names and faces (Bastin, Van der Linden, Schnakers, Montaldi, & Mayes, 2010). These three types of item pairings (within-item, intra-domain, and inter-domain, respectively) for associative memory are the traditional pairings, but they are not exhaustive - there are other ways in which information can be bound together. One example is list retrieval tasks, where participants study two separate lists of words, and at test must indicate whether the target word was originally presented in the first or second list (Whiffen & Karpicke, 2017). Another example is the context effect on memory retrieval: Memories are better retrieved if the studied object is presented in the same background or context as at encoding (Oliva & Torralba, 2007), and this extends to words, faces, and nonwords (Russo, Ward, Geurts, & Scheres, 1999). A full review of different tests of associative memory is provided by Mayes, Montaldi, and Migo (2007).

Associative memory is often tested with an Old-New task or a 2AFC task. A key difference between the construction of Old and New pairs in associative memory research, and other recognition research, is that the 'New' pairs in the associative test *do not* comprise New items. Instead, 'New' pairs comprise Old items that were not paired with each other at encoding, whereas 'Old' items comprise a pair of items that were shown together at encoding. This is to safeguard against participants distinguishing between pairs because they recognise one of the items within that pair as New instead of recognising that the pairing is incorrect.

There is a great deal of research on paired-associates for within-domain items, especially for word-pairs (e.g., Becker, 1979; Krause et al., 1999; Lowndes et al., 2008; Perea & Rosa, 2002; Putnam, Ozubko, Macleod, & Roediger, 2014), and some research on face-pairs (Bastin & Van der Linden, 2005; Naveh-Benjamin, Guez, Kilb, & Reedy, 2004; Rhodes, Castel, & Jacoby, 2008; Watkins, Ho, & Tulving, 1976; Winograd, & Rivers-Bulkeley, 1977). Asking eyewitnesses *who* did *what is* an associative memory test for between-domain items. Most of the research that used between-domain pairs with faces used face-name pairs (Bender et al., 2017; James, Fogler, & Tauber, 2008; Naveh-Benjamin et al., 2004; Naveh-Benjamin et al., 2009), although there has been research that used face-spatial locations (Bastin & Van Der Linden, 2005; Crook, Larrabee, & Youngjohn, 1993[55]).

Findings from Bastin et al. (2010) suggest that associations for between-domain items are more poorly recognised than associations for within-domain items. Bastin and colleagues instructed their participants to study two sets of stimuli: one set comprised 40 face-face pairs, and the second set comprised 40 face-name pairs. Following encoding, participants were tested on the pairs, and individual components that comprised the pairs. Recognition was tested using 2AFC tests; for the 2AFC test for pairs, participants were shown a pair of items (face-name, or face-face) that were either shown together at encoding, or were not shown together at encoding. Bastin et al. (2010) also manipulated the recognition instructions between participants, so that half were instructed to indicate which of the two items presented in the 2AFC test they recognised, whereas the other half were instructed to indicate which item was familiar. The

[55] Spatial locations were learnt incidentally. Participants were instructed to study face and name pairs, and the same faces were presented with different spatial locations. This is different to other paired associate tests where participants are tested on the two associated pieces of information shown at encoding.

emphasis on familiarity stems from dual process models of memory (Atkinson & Juola, 1974; Jacoby, 1991; Mandler, 1991; Tulving, 1985a; Tulving, 1985b; Yonelinas, 1994), which posit that two processes, familiarity and recollection, underlie memory. Familiarity refers to recognising an item as previously studied without the simultaneous recollection of the encoding context; recollection refers to recognising an item as previously studied accompanied by the conscious recollection of the encoding context. For this reason, familiarity relies on a more relaxed criterion than recollection.

The results from Bastin et al. (2010) are presented in Table 3.5. The authors analysed face recognition, name recognition, and pair recognition separately. The only significant result for face recognition was a main effect for instruction type on accuracy. Overall, participants who received the familiarity instructions performed worse on the face recognition tasks than those who received the standard instructions. For participants who were tested on names, there was also a significant difference in accuracy between the two instructions groups. Participants who received the standard instructions performed better on the name recognition tasks than participants who received the familiarity instructions. For the associations test, there was a significant main effect for instruction type, and a significant interaction between instruction type and test type. Participants performed worse at the pairs test following the familiarity instructions than the standard instructions. When instructions by test type was compared, there were only two significant pairwise differences: Within the face-pair group, there was a significant difference between standard and familiarity instructions, and within the standard instructions group, there was a significant difference between test types. Face-name pairs were better recognised than face-face pairs.

Table 3.5

Results for Face-Face Pairs and Face-Names Pairs as Presented Reported in Bastin et al., 2010.

				Proportion Correct	
Encoding	Number of Trials at Encoding	Test Type	Number of Trials at Recognition	Standard Instructions	Familiarity instructions
Face-face pairs	40	Face recognition	20 (4 practice trials; 16 test trials)	.92	.71
		Face-face pairs	20 (4 practice trials; 16 test trials)	.70	.65
Face-name pairs	40	Face recognition	10 (2 practice trials; 8 test trials)	.95	.72
		Name recognition	10 (2 practice trials; 8 test trials)	.91	.75
		Face-pair recognition	20 (4 practice trials; 16 test trials)	.82	.57

Note. These table is reconstructed with the values from Table 1 and Table 2 reported in Bastin et al., 2010.

Within the between-domain pair associations literature, specifically the face-name pair associations literature, there is a small pocket of research that has investigated recognition memory for face-name pairs, as well as memory for the individual items that comprise these pairs (Bastin et al., 2010; Bender et al., 2017;[56] Naveh-Benjamin et al., 2004; Naveh-Benjamin et al., 2009). The study characteristics and results from the three studies are in Table 3.6.

[56] Many of the studies reported by Naveh-Benjamin investigate memory impairment of older adults, specifically the marked decline in associative memory. Naveh-Benjamin has repeatedly shown that the decline is associative memory is not due to an overall decline in memory performance, since older adults can remember the individual components that comprise the associative memory. See Naveh-Benjamin (2000) for a discussion of the age-related impairment in associative memory.

Table 3.6

Hit Rates and False Alarm Rates for Studies that Tested for Item Memory, and Associative Memory Following Encoding of Face-Name Pairs

		Encoding	Test			Results	
Study	Participants	Set Size	Format	Type	Set Size	HR	FAR
Bastin et al., 2010, Standard Instructions	Adults	40	2AFC	Name	8	.91[b]	-
				Face	8	.95[b]	-
				Pairs	16	.82[b]	-
Bastin et al., 2010, Familiarity Instructions	Adults	40	2AFC	Name	8	.75[b]	-
				Face	8	.72[b]	-
				Pairs	16	.57[b]	-
Bender et al., 2017, Exp. 1	Young adults	48	2AFC	Name	12	≈.68[a,b]	-
				Face	12	≈.65[a,b]	-
				Pairs	24	≈.68[a,b]	-
Bender et al., 2017, Exp. 2a	Young adults	32	2AFC	Name	8	.81[b,c]	-
				Face	8		-
				Pairs	16	.72 [b]	-
	Older adults	32	2AFC	Name	8	.79[b,c]	-
				Face	8		-
				Pairs	16	.66 [b]	-
Naveh-Benjamin et al., 2004	Young adults	40	2AFC	Name	16	≈.80[a,b]	-
				Face	16	≈ .92[a,b]	-
				Pairs	16	≈ .73[a,b]	-
Naveh-Benjamin et al., 2009; Exp. 1	Young adults	27	Old New	Name	16	.69	.16
				Face	16	.72	.13
				Pairs	unknown[d]	.68	.30
	Older adults	27	Old New	Name	16	.69	.16
				Face	16	.78	.15
				Pairs	unknown [d]	.72	.45
Naveh-Benjamin et al., 2009; Exp. 2	Young adults	48	Old New	Name	32	.72	.16
				Face	32	.76	.22
				Pairs	32	.70	.31
	Older adults	48	Old New	Name	32	.54	.24
				Face	32	.76	.42
				Pairs	32	.61	.49

Notes. HR and FAR denote Hit Rate and False Alarm Rate respectively. All the studies presented Face-Name Pairs at encoding.
[a] No descriptive statistics were provided, so these values were estimates from the figures.
[b] These values are not true 'Hits', but actually represent % Accuracy.
[c] These studies did not report the mean accuracy for Name and Faces respectively, but instead reported a mean accuracy averaged across both tests.
[d] These studies did not report the number of trials completed for the Pairs test.

Inspection of Table 3.6 shows that set size at encoding ranged from 27 to 48, and only a subset of encoded items was tested at recognition. Across all experiments, recognition performance for item-tests was superior to recognition memory for pair-tests. For experiments that used Old-New tests, FAR was highest for pairs compared to items. It was not possible to conclude whether recognition memory was better for some types of items than others: For some tests, performance was highest for names, whereas for others, performance was highest for faces.

Most of the experiments in Table 3.6 used a 2AFC test at recognition. Recognition performance was generally higher following the 2AFC tests than Old-New tests. This could be attributed to task difficulty: Old-New tasks are more difficult than 2AFC tasks. For a 2AFC task participants can use two strategies to make a decision: Participants can rely on the presence of familiarity surrounding one stimulus to choose that stimulus or on the absence of familiarity from one stimulus to choose the other stimulus. For Old-New tasks, the participant can only respond to the currently presented stimulus. Overall, it is recognised that 2AFC tasks lead to higher recognition performance than Old-new tasks (Jang, Wixted, & Huber, 2009).

None of the studies in Table 3.6 controlled for picture recognition between encoding and recognition of faces. The high accuracy could be partially accounted for by the lack of variation between stimuli at encoding and recognition (Bruce, 1982). Within the associative memory literature reviewed thus far, controlling for picture recognition has not been mentioned, and it remains unclear how to do this for paired memory tasks. However, researchers of associative memory may not want to introduce variation in the images used at encoding and recognition, as this might reduce the strength of the paired information at test.

Of the studies listed in Table 3.6., only one experiment manipulated set size at encoding (Experiment 2a, Bender et al., 2017). In their experiment, two sets of participants (older adults and younger adults) completed three encoding trials, where they studied a different set size of face-names pairs in each trial (32, 40, or 48 pairs). After encoding and a 60 second delay interval, participants were tested on faces, names, and face-name pairs using 2AFC tasks. The authors averaged the mean accuracy performance for both faces and names – called item accuracy. The complete descriptive statistics are not reported for this experiment, so it is not possible to compare the recognition results between faces and names for set size. The authors only reported that there was a significant interaction between set size and test type in absence of a main effect for either of these two factors: There was no significant difference between associations and items for Set Sizes 40 or 48, but recognition accuracy was better for associations than items at Set Size 32. However, without any descriptive statistics or accompanying graphs, it is difficult to interpret these results.

Of the research on between-domain pairs, only one study looked at the associations between people and activities (Old & Naveh-Benjamin, 2008), which could be a precursor for eyewitness research on perpetrator-role pairings. Old and Naveh-Benjamin (2008) recruited older and younger adults and showed them 75 video clips, each depicting a unique individual performing a unique action (e.g., bouncing a ball). Each video was five seconds long. Following encoding and a five-minute filler task, participants' recognition memory was tested using 2AFC tasks for (a) the actor in the video, (b) the action performed in the video, and (c) the actor-action pair. For the actor-test, participants were shown video clips of the actor sitting idly. For action-test, participants were shown video clips of the hands of an actor performing the action; the face of the actor was not visible. For the actor-action pairs, participants were shown video clips of the actor performing the same action as shown at encoding (i.e., a congruent pair), or a previously seen actor performing a previously seen action, but not the action that they originally performed in the encoding video (i.e., an incongruent pair). Participants were only tested on a subset of the pairs: The 2AFC task for actors and actions comprised 14 clips each (7 old, 7 new), whereas the 2AFC task for actor-action pairs comprised 18 clips (9 old, 9 new). Following these three tests, participants returned the next day to complete a second set of tests (which used different items) to determine what impact delay had on recognition performance. The results are presented in Table 3.7. Overall, there was a significant main effect for the delay period, and the type of test. Performance was better following a short delay versus a long delay, and was best for actions, followed by associations, and then faces.[57] There were no significant interactions.

Table 3.7

Hit Rate and False Alarm Rate for Actions, Faces, and Pairs After Two Delays as Reported in Old and Naveh-Benjamin, 2008

	After 5-minute delay						After 24-hour delay					
	Actions		Faces		Pairs		Actions		Faces		Pairs	
Sample	HR	FAR	HR	FAR	HR	FAR	HR	FAR	HR	FAR	HR	FAR
Younger Adults	.87	.03	.72	.18	.91	.18	.74	.06	.53	.18	.82	.36
Older Adults	.78	.04	.68	.26	.86	.38	.63	.08	.52	.27	.75	.52

Notes. HR denotes Hit Rate, and FAR denotes False Alarm Rate

The authors interpreted their results within the context of eyewitness memory suggesting that older adults may struggle to recall perpetrator-actions, that is, correctly pairing the perpetrator to the roles (see paragraph 1 of the Discussion in Old & Naveh-Benjamin, 2008). The authors may be correct - older adults may, indeed, struggle to recall the finer details of the crime – but it is unlikely that an eyewitness/victim to a single-perpetrator crime would forget that the perpetrator was broadly responsible for the crime (and all the actions that it entails); I have made this argument throughout this thesis. Instead, it is unclear how eyewitnesses to multiple-perpetrator crimes would fare at a perpetrator-role pairing task. Old and Naveh-Benjamin (2008) had shown their participants multiple actor-action pairs at encoding (in total, 75 videos), but they were tested on only a subset of the materials. Thus, the recognition performance reported in Old and Naveh-Benjamin may be an overestimate of recognition ability.

[57] Old and Naveh-Benjamin (2008) calculated the difference between the proportion of hits and the proportion of false alarms. They termed this difference ‘performance’ and all analyses were conducted on this newly calculated value.

The Role of Familiarity and Recollection in Associative Memory Tasks

The studies reviewed in the current chapter used either 2AFC or Old-New tasks to test for memory, and the stimuli shown at test for paired associations were either two items that were shown together at encoding (i.e., two correctly matched Old items), or two items that were shown at encoding, but not together (i.e., two incorrectly matched Old items). Old-New and 2AFC tests are the two primary methods used to test associative memory (Mayes et al., 2007; Old & Naveh-Benjamin, 2008; Yonelinas, 2002), but there are other ways to test associative memory. For example, participants might be presented with only one component from the originally encoded pair, and must respond with the component that will complete the pair. In this example of a cued-response task, the presented component to which participants respond (i.e., the Old item) acts as a cue for the component that will complete the pair.

The sub-processes of familiarity and recollection, as outlined in dual-process models of memory (Jacoby, 1991; Tulving, 1985a; Tulving, 1985b; see Yonelinas, 2002, for a thorough review of dual-process models), play an important part in associative memory. In tests of associative memory, item-pairs that consist of only old items are meant to prevent participants from relying on the familiarity sub-process. Since both items within the pair were presented previously ay encoding, participants cannot merely respond to the item that feels the most familiar. Instead, participants must respond to the pair using the recollection sub-process, that is, participants must actively recall whether the items were presented with each other.

Metacognitive strategies surrounding familiarity and recollection are often measured using self-introspect questions. For example, Bastin et al. (2010) gave their participants either 40 face-face pairs or face-name pairs to encode, and at test, instructed participants to either indicate which pair they recognised, or which pair was familiar. Participants who received the ‘familiarity’ instructions were told to make a decision based on the ‘first experience/awareness of familiarity’. Tulving (1985b) asked his participants to provide a Know-Remember response about the metacognitive processes that accompanied their recollection for memory response. Participants were instructed to make a Know response when they had a sense of familiarity towards the test stimulus *and* could not recall any of the encoding context, and a Remember response when they could recall episodic information about the encoding context surrounding the test stimulus. Thus, Know responses measure familiarity, and Remember responses measure recollection.

There are several ways to compare memory performance for Know and Remember responses. The two sub-processes are considered mutually exclusive, and the proportion of Remember and Know responses can be compared to determine which memory process contributed the most to the memory test. A second way to compare Remember and Know responses is to plot Hits and False Alarms using Receiver Operator Characteristic (ROC) curves (Yonelinas, 2002; Yonelinas & Parks, 2007). Normally Hits and False Alarms are plotted at various decision criteria, such as confidence levels. For Know-Remember experiments, participants provide a binary Remember-Know response to recognised stimuli, followed by a confidence rating on (for example) a three-point scale. Hits and False Alarms can be arranged along this six-point confidence scale, with the three confidence points for Know and the three confidence points for Remember comprising the lower and upper bands of the confidence scale respectively. Then ROC curves are plotted for the Hits and False Alarms for the six confidence points.

Eyewitness memory for associative memory may behave differently to associative memory measured in the laboratory. For example, unlike cued-response laboratory tasks where an Old item acts as a cue, eyewitnesses may attempt to recall the role for an *innocent* suspect whom they *mistakenly* identified. This would have little effect on the ability of an eyewitness to a single-perpetrator crime to recall the associated role, because there is only one possible role to recall. For eyewitnesses to multiple-perpetrator crimes, however, role-recollection following a mistaken identification from a target-absent parade would be exceptionally difficult. The innocent suspect, who was misidentified, cannot cue any associated role. Thus, there is no memory association between the suspect and the role, and cueing should fail. In such situations, eyewitnesses can only make their decisions based on familiarity (instead of true recollection), or by using a process of elimination/degrees-of-freedom strategy where eyewitnesses may choose the role that is ‘left over’ after assigning other roles to previously identified suspects. Thus, it is possible that set size may impact the metacognitive strategies used by participants when responding to associative memory tasks.

Conclusion

The literature reviewed in the current chapter demonstrated that human memory can store large amounts of visual data, but estimates of memory capacity for visual items may not include memory for faces. While it is recognised that memory for familiar faces is robust, human memory for unfamiliar faces is less resilient. Only three studies manipulated set size at encoding for unfamiliar faces (Lamont et al., 2005; Metzger, 2002; Podd, 1990), and while all three found that set size impaired discriminability, each found a different pattern of Hits and False Alarms. None of these studies controlled for picture recognition by varying the stimulus materials between encoding and recognition; as a result, these recognition results may not be a true reflection of the ability to recognise a facial identity.

The reviewed literature also demonstrated that item memory is better retained than associative memory. Only one reviewed study manipulated set size for item pairs presented at encoding, but the results are difficult to interpret because no descriptive statistics were reported (Experiment 2b, Bender et al., 2017). Thus, it is unclear whether item memory and associative memory are immune to effects of set size, or whether set size has a detrimental effect on recognition performance for both types of memory tests.

Furthermore, associative memory is meant to rely on the recollection sub-process, rather than familiarity sub-process; however, this depends on the type of test used to test associative memory. If participants are asked to respond to pair of Old items, then they are more likely to rely on recollection sub-processes. If, however, participants are shown only one Old item, and are asked to provide the matching item that completes the pair, then they are less likely to only rely on recollection sub-processes. Furthermore, the relationship between dual-process models of memory and set size is unclear. It is possible that recollection sub-processes are used less frequently as set size increases, that is, participants may be more likely to rely on familiarity sub-processes at test after encoding large sets of paired items.

Chapter 4

Testing the Effect of Set Size on Face Recognition and Associative Memory

Aim and Rationale

In this chapter I present two face recognition experiments that test the effect of set size on recognition memory for faces, attributes, and associative memory. The aim and rationale for these two experiments follow from the findings in Chapter 2, and the literature in Chapter 3. From the police survey presented in Chapter 2, we know that (a) eyewitnesses are required to qualify their identifications with supporting information, and (b) eyewitnesses to multiple-perpetrator crimes may be required to identify more than one perpetrator. It remains unclear, however, whether eyewitnesses can correctly identify multiple perpetrators.

In Chapter 3, I reviewed the face recognition and associative memory literatures. Podd (1990), Metzger (2002), and Lamont et al. (2005) investigated the effect of set size on face recognition and found that recognition performance decreased as set size increased; all three studies, however, only tested recognition memory for a subset of the encoded stimuli. Furthermore, there were mixed results concerning the reason for the decreased recognition performance: Was the decreased recognition performance due to a decrease in Hits, an increase in False Alarms, or both?

Research on associative memory has reliably demonstrated that item memory is better than associative memory; however only one study tested the effect that set size at encoding has on associative memory (Experiment 2b, Bender et al., 2017). Despite the authors' conclusion that set size made no impact on recognition memory, the readers cannot judge this themselves, since descriptive statistics were not reported. Additionally, among the studies reviewed, recognition memory for associative memory was frequently tested with a 2AFC test, which is considered an easier memory test than an Old-New Test (Jang et al., 2009), and may yield better recognition results.

Thus, the effect of set size on face recognition memory and associative memory remains unanswered. The aims of the two experiments in this chapter are to determine how set size affects (a) face recognition, and (b) associative memory recognition. Both experiments use the same method and the same materials, but a different recognition test to measure associative memory. In both experiments, participants studied a set number of face-attribute pairs, and were tested on item memory (faces, attributes) using an Old-New test. In Experiment 1, associative memory was tested with a cued-matching task: Participants were shown old faces, one at time, and given a list of old attributes, and had to decide which attribute from the list matched the face presented on the screen. In Experiment 2, associative memory was tested with an Old-New task where participants were shown congruent pairs (old faces and old attributes that were studied together) or incongruent pairs (old faces and old attributes that were shown at encoding, but not paired together).

Additionally, in the eyewitness literature, confidence is normally considered a postdictor of accuracy (Robinson, Johnson, & Herndon, 1997; Sporer, Penrod, Read, & Cutler, 1995), or an initial measurement of confidence is used as a predictor of accuracy (Wixted, Mickes, Clark, Gronlund, & Roediger, 2015). Confidence, however, is not a stable measure that is only related to accuracy; instead, it can also change because of repeated lineup decisions (Godfrey & Clark, 2010), post event questioning (Shaw, 1996; Shaw & McClure, 1996), and lineup format (Dobolyi & Dodson, 2013). In fact, in some situations, confidence can be an extremely unreliable measure, as has been found in post-identification feedback experiments (Wells & Bradfield, 1998; Tsunga, 2018). In the post-identification feedback paradigm, participants are forced to make a decision from a target-absent parade, and upon receiving positive feedback about their incorrect lineup decision, they report higher confidence in their lineup choice. Dobolyi and Dodson (2013) argue that sequential parades induced a more conservative choosing criterion than simultaneous parades, and, consequently, identifications for sequential parades are made with higher confidence. This is because a conservative criterion leads to decision making that is only accompanied by a strong memory trace, and a strong memory trace is accompanied by higher confidence. Thus, eyewitnesses report higher confidence for decisions made from sequential parades than simultaneous parades due to the conservative criterion that is induced by the parade format. This highlights the complexity of measuring confidence – it is not a pure measure of the eyewitness' belief in their decision-making abilities, since it can be inflated under unusual situations (like the post-identification feedback scenario), and it is confounded by task difficulty, lineup format, postevent questioning, and criterion. In the set of experiments reported in this chapter, confidence is treated as an outcome variable, since set size might impair confidence. Each experiment, and its results, is discussed in turn below.

Face Recognition Experiment 1:
Testing Memory for Pairs Using a Cued-Matching Task

Method

Design

This experiment contained one between-subject independent variable, set size, with seven levels, and one within-subject variable, test type, with three levels. Participants were randomly assigned to one of seven levels of set size, where the number of face-attribute pairs that they studied varied within each trial (one, two, three, five, ten, fifteen, or thirty face-attribute pairs). The dependent variables, accuracy and confidence, were measured for each of the three levels of test type: faces, attributes and face-attribute pairs.

Sample

Seventy students (*Mean age* = 20.02 years, *SD* = 2.28; 56 females) were recruited from the undergraduate student body through the Student Research Participation Programme (SRPP).

Materials

Faces at encoding. Two sets of encoding materials (faces, attributes) were generated. Thirty synthetic faces of young men were generated from bespoke software, ID (Tredoux, Nunez, Oxtoby, & Prag, 2006). Synthetic faces, rather than real faces, were used to ensure a large enough sample of images where similarity was controlled for between target and distractor images. To create these artificial images, the total sample of photographs of coloured[58] South African males were selected from a database of photographs curated by Tredoux, and reduced to 110 images by removing older adults and highly distinctive faces that would introduce model artefacts. This reduced sample was entered into ID, which isolated the shapes and textures from the photographs to form a vector space that represented the characteristics of sample images. To do this, ID uses principal component analysis to generate eigenfaces, which form the face space. From the vector space, ID can generate synthetic faces, but with a high degree of realism and which are not replicas of the original sample from which the space was constructed (although these synthetic faces are bound by the characteristics of the original sample). Synthetic faces are generated from eigenface coordinates within the vector space; these eigenface coordinates do not map to the original faces used to construct the space and can exist anywhere within the vector space. For this reason, ID can generate a greater number of synthetic faces than what was used to build the vector space originally.

[58] In South Africa, the demographic term 'coloured' dates to the Apartheid era, and was used to define people who were of 'mixed race'. This term is still used today, but the term is not limited to physical appearance (although it was used in this manner originally). This group of individuals, whom the term defines, is recognised as a group with its own cultural identity.

The 30 synthetic faces generated from ID were in colour, frontal-view, with neutral expressions (see Figure 4.1). Using Adobe Photoshop CS3, all images were digitally edited to remove pictorial artefacts, a t-shirt was superimposed at the neckline, and the entire image was superimposed onto a dark grey, wall with a stucco texture (see Figure 4.1). The images were standardised for interocular distance.

Figure 4.1. Two versions of a target face used at encoding for Face Recognition Experiment 1. The image on the left is the original image generated by ID. The image on the right is the edited version of the same image that was used at encoding.

Attributes at encoding. Initially, I generated a list of fifty unique attributes, which each described a hobby, characteristic, or preferences, for example, ‘He brews his own beer’. With the assistance of a second, independent rater, I refined the list by removing any attributes with strong valence, for example political behaviour (e.g., ‘He voted for the DA’). The final list was reduced to thirty attributes (Appendix F). Each attribute had a length of approximately ten words.

Face-attribute pairs at encoding. The pairing between faces and attributes was randomly generated prior to the experiment, and two sequences of pairings (and faces, and attributes) were constructed to control for any sequence effects. Even though E-Prime (Psychology Software Tools, Pittsburgh, PA; the software used to deliver the experiment) can generate a random pairing of faces and attributes within the experiment, I preferred two-predetermined lists because of the procedure utilised in the recognition stage in this experiment (this will be explained in the procedure section). This random pairing, and the two sequences, are presented in Appendix G.

Faces at recognition. Two additional sets of materials (faces, attributes) were generated for the recognition phase. Thirty new faces were generated using a feature of ID where the user indicates the desired similarity (i.e., relative proximity within the face space as measured in Euclidean distance) between a new face and a selected face. Thus, a new face was generated for each of the old faces using a DIFS value of 0.09 (DIFS denotes distance within face space – this is a Euclidean distance); various DIFS values were sampled and this value (0.09) yielded images of high subjective similarity (see Figure 4.2.). These new faces were edited in the same manner as the old faces: Pictorial artefacts were removed, the grey, stucco background was inserted behind the face, and a white t-shirt was superimposed at the neckline.

A current limitation of ID is that it can only generate frontal-view images; thus, only one version of each old and new face was generated. This was problematic, because there was a risk that participants would recognise the picture from an artefact of that version of the image, instead of recognising the face of the individual (Bruce, 1982; Bruce et al., 1999). To facilitate facial recognition while working within the constraints of ID, the old and new faces were transformed from colour to greyscale images (Figure 4.2). There are various ways to buttress against picture recognition, but most of these are dependent on sample availability, for example, having multiple photographs of the target image. Since I did not have a larger sample available, I tried to induce variation between the encoding and recognition stimuli by using a colour version at encoding, and a greyscale version at recognition.[59]

Old Image **New Matched Image**

Figure 4.2. An example of an old face and new face used in this experiment. The image on the left is of an old face that was transformed to greyscale for the recognition stage. The image on the right is of its matched distractor face (i.e., a new face) that was also used for the recognition stage.

Attributes at recognition. Thirty new attributes were generated from the old attributes. The two sets of old and new attributes differed only on the main characteristic of each attribute, for example, the old attribute 'He hates raisins' was matched with the new attribute 'He hates prunes'. The new attributes were generated with the assistance of a second rater. To further introduce variation in the attributes between encoding and recognition, attributes were presented in a different font at the two stages in the experiment A list of old and new attributes are presented in Appendix F.

[59] Other options are to flip the image on the vertical axis, or to pixelate the image.

Procedure

Participants entered the laboratory, signed a consent form and were seated in front of individual computers where they were randomly assigned to one of seven experimental groups. The experimental groups differed according to the number (i.e., set size) of face-attribute pairs shown at encoding (one, two, three, five, ten, fifteen or thirty pairs). The experiment was delivered using E-Prime 2.0.8 (Psychology Software Tools, Pittsburgh, PA).

The experiment was divided into three stages: encoding, filler, and recognition. At the start of the encoding stage, participants were instructed that they would see a series of face and attribute pairs, presented one pair at a time, which they had to study together because they would be tested at a later stage. Participants in the Set Size 1 group were given a variation of these instructions: They were told that they would view one face and attribute pair (instead of a series). After these instructions, a small fixation cross appeared on the screen for 250ms, followed by a target face that appeared for three seconds. After three seconds, an attribute was presented simultaneously with the face for an additional three seconds. Following this, the face disappeared whereas the attribute remained on the screen for a further three seconds before disappearing. Total encoding time was six seconds for the face and the attribute respectively, and the face and the attribute appeared together simultaneously on the screen for three seconds. All participants - except those in the Set Size 1 condition - repeated this encoding procedure until all face-attribute pairs for their experimental group were presented. As mentioned, the pairing of faces and attributes was pre-determined (see Appendix G), but the order of these pairings within each encoding stage was random. An inter-stimulus interval (ISI) of 250ms, followed by a fixation cross presented for 250 ms, appeared between each face- attribute pair.

After encoding, participants engaged in an unrelated filler task. The time assigned for the filler task varied among the experimental groups to account for the differences in the delay between the presentation of the median face within the set at encoding and the start of the recognition trial. Specifically, participants in the Set Size 30 group had the longest delay between the median face in the encoding set and the start of the recognition stage, whereas participants in the Set Size 1 group had the shortest delay between encoding and recognition. The duration of the filler task for the Set Size 1 and Set Size 30 groups was 137.75 seconds and 60 seconds, respectively, whereas the duration of the filler task for the remaining groups varied between these two values so that the delay between the presentation of the median face at encoding and recognition was held constant.

Following the filler task, participants began the recognition stage, which was divided into three different tests. First, participants completed an Old-New task for faces. Participants were shown old faces and new faces, one at a time in a randomised order, and had to indicate by keypress whether they had previously studied the presented face. After making an Old-New decision, participants were asked to rate their confidence in their decision on a scale from 0% to 100%. After participants had responded to all old and new faces, they began the Old-New attribute recognition test. Participants were shown old and new attributes, one at a time in a randomised order, and had to indicate by keypress if the presented attribute was old or new. Afterwards, they rated their confidence in their decision. After participants had responded to all the attributes, they were presented with a third test – a cued-matching task - where they had to match old attributes to old faces. Participants were a given a printed list of all the attributes that they had studied within that trial, saw each of the old faces that they encoded during that trial, one at a time in a randomised order on the screen, and had to decide which attribute from the list was originally paired with the presented face. Participants could only move forward to the next face once they had responded to the current face. Attributes could be paired multiple times if participants were uncertain. No new faces or attributes were shown during the pairing test. Participants in the Set Size 1 group received a list that contained only one attribute, and saw only one face during this task; thus, it was expected that they would achieve 100%

accuracy for this task. After each pairing decision, participants were asked to rate their confidence on a scale from 0% to 100%.

This procedure – encoding, filler task, and recognition – was repeated until participants had studied all thirty faces (see Appendix H). Therefore, participants in the Set Size 1 group completed thirty trials of this procedure, and were presented with a different face-attribute pair in each trial. In contrast, participants in the Set Size 30 group completed only one trial of this procedure, but they studied all thirty face-attributes pairs at once. Thus, across all conditions, all participants studied thirty face-attribute pairs, but did so under varying set size loads (see Appendix H). Participants were debriefed once the entire experiment was completed. The procedure is shown in Figure 4.3

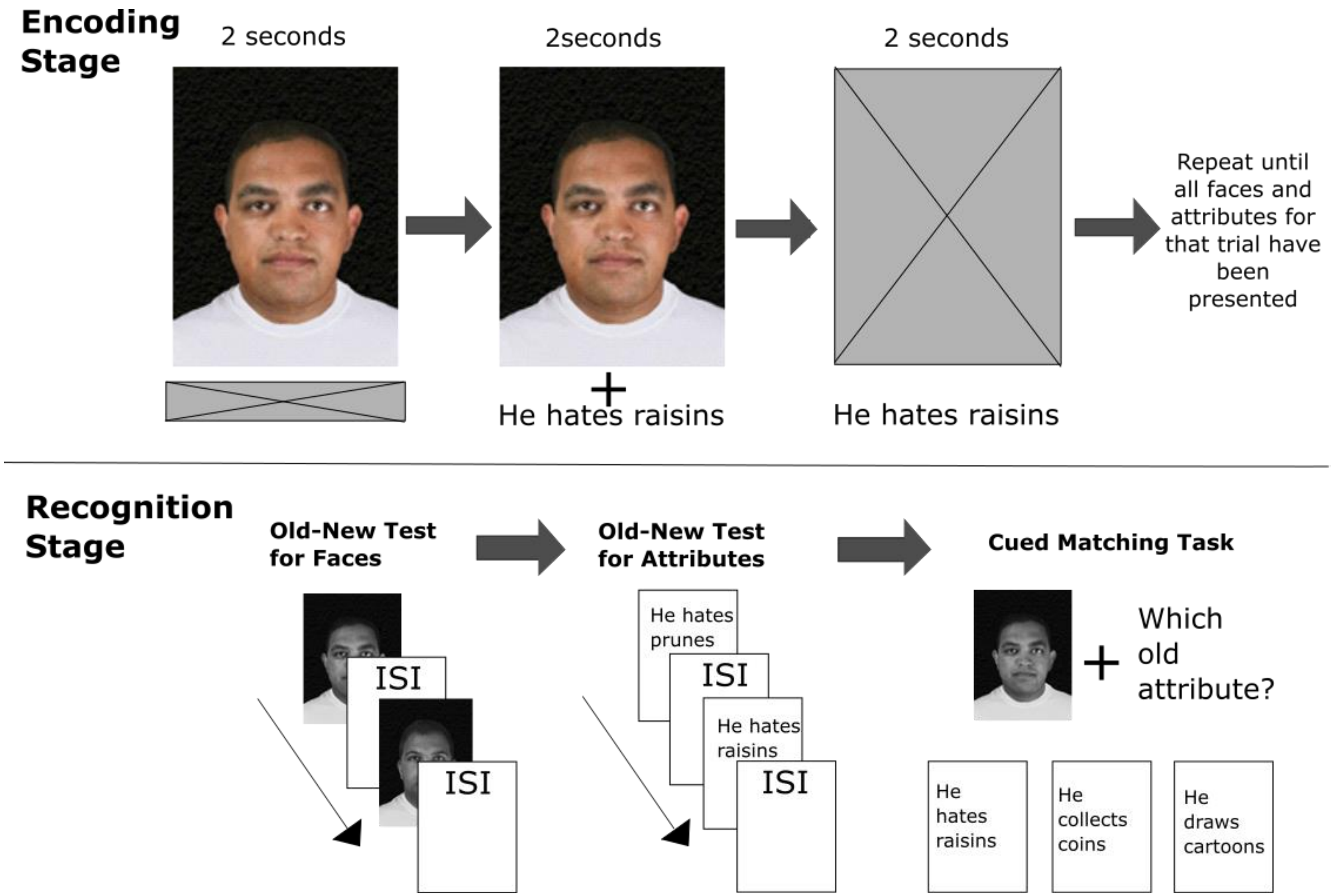

Figure 4.3. Illustration of the procedure used in Face Recognition Experiment 1. During the encoding stage participants were presented with face-attribute pairs for a total of nine seconds, and after a short filler task, completed the recognition stage. During the recognition stage, participants completed three tests: an Old-New test for faces, an Old-New test for attributes, and a cued matching task where they were given a list of all the old attributes in that trial and had to decide which of these matched the old face presented on the screen.

Results

Participants responded to three tests: a face recognition test, an attribute recognition test, and a cued matching task. Hit Rate (HR) and False Alarm Rate (FAR) were calculated for each participant for each test.

Discriminability and Criterion

From these two statistics (HR, FAR), it is possible to calculate discriminability and criterion (Macmillan & Creelman, 2005). Discriminability, typically abbreviated as *d'*, refers to how well an individual can distinguish between Hits and False Alarms, which in a face recognition experiment is identical to a participant discriminating between old and new items. Discriminability can improve for three reasons: HR improves (i.e., participants respond correctly to more old items), FAR decreases (i.e., participants respond correctly to more new items), or a combination of both. The relationship between HR and FAR is important: If only HR is studied (while ignoring FAR), then the researcher could incorrectly conclude that the participant is good at identifying old items when, instead, the participants responds affirmatively to both old and new items (i.e., indicating that they recognise both items as 'old'). This response tendency is called criterion.

HR and FAR can be analysed separately (so long as FAR is not ignored), or can be reduced to a single statistic, *d'*. Discriminability, or *d'*, is calculated from the following formula *d'= z(proportion of Hits) – z(proportion of False Alarms).* It is not possible to calculate *z* when HR or FAR is extreme (e.g., 100% or 0%; since z approaches ∞ asymptotically) without implementing one of two possible solutions described in Macmillan and Creelman (2005). [60] The middle point for *d'* is zero, with higher values indicating better discriminability.[61]

Criterion is calculated with the following formula, *c =-0.5(z[proportion of Hits]+z[proportion of False Alarms]).* The centre point for *c* is zero, and a negative and positive criterion indicate lenient and conservative bias respectively. That is, negative values indicate a willingness to respond 'old', whereas positive values indicate a willingness to respond 'new'.

While it is possible to calculate *d'* and criterion for the results for the faces and attributes tests respectively, *d'* cannot be calculated for the cued-matching task. The cued-matching task was a forced-choice test: Participants were shown only old faces, were given a list of old attributes, and had to decide which item from the list of old attributes was paired with the presented old face. Forced-choice decisions do not require a threshold for deciding between 'old' and 'new', nor are there any true 'false alarms' (instead, wrong decisions are errors). Without a threshold, and without a distribution of false alarms underlying the responses, it is not possible to calculate *d'*. Additionally, there is little value in calculating *d'* for only two of the tests since the aim of this experiment is to compare performance across the three types of tests. Thus, only HR and FAR were calculated for this first face recognition experiment; *d'* is, however, calculated for the second face recognition experiment, hence these concepts are introduced now.

---

[60] There are two recommended methods to correct extreme values. First, perfect scores of 1, or scores of 0 can be corrected by 1 - 1/(2N) and l/(2N) respectively. N refers to the number of trials for that statistic, that is, N is the number of trials from which the Hit is calculated. For the second method, 0.5 is added to the total number of all cells (Macmillan & Creelman, 2005; Stanislaw & Todorov, 1999).

[61] *d'* can also be negative, implying more false alarms than hits. Stanislaw and Todorov attribute this to either instruction confusion or to sampling error (see page 139 of Stanislaw & Todorov, 1999).

Hits and False Alarms for Faces, Attributes, and Pairs Across Set Size

HR and FAR were calculated for each of the three tests for the different set sizes and are reported in Table 4.1 HR for each of the three tests decreased as set size increased, but of the three tests, HR decreased the least for attributes, and decreased the most for pairs. In contrast, FAR increased drastically for pairs, increased and then stabilised for faces, and increased minimally for attributes.

Table 4.1

Hit Rate and False Alarm Rate Across Test Type and Set Size

	Faces		Attributes		Pairs	
Set Size	HR	FAR	HR	FAR	HR	FAR
1	0.94 (0.02)	0.07 (0.02)	0.99 (0.01)	0.01 (0.01)	0.99 (0.00)	0.01 (0.00)
2	0.95 (0.02)	0.12 (0.03)	0.99 (0.00)	0.01 (0.01)	0.96 (0.02)	0.04 (0.02)
3	0.93 (0.01)	0.15 (0.04)	0.98 (0.01)	0.03 (0.01)	0.87 (0.04)	0.13 (0.04)
5	0.87 (0.03)	0.19 (0.04)	0.97 (0.01)	0.03 (0.01)	0.69 (0.07)	0.31 (0.07)
10	0.81 (0.04)	0.23 (0.04)	0.96 (0.01)	0.07 (0.02)	0.32 (0.03)	0.68 (0.03)
15	0.76 (0.03)	0.26 (0.05)	0.95 (0.02)	0.05 (0.02)	0.27 (0.04)	0.73 (0.04)
30	0.64 (0.05)	0.21 (0.04)	0.91 (0.02)	0.07 (0.03)	0.08 (0.02)	0.92 (0.02)
Average	0.84 (0.02)	0.18 (0.02)	0.97 (0.01)	0.04 (0.01)	0.60 (0.04)	0.40 (0.04)

Note. HR and FAR denotes Hit Rate and False Alarm Rate respectively. There were thirty trials for faces, attributes, and pairs. Values in parentheses denote standard errors. The final row denotes the overall average across set size for the three types of test.

The mean HR and FAR were plotted (Figure 4.4), and the pattern of results suggested that (a) recognition memory for attributes was resilient and remained somewhat unchanged across set size, (b) recognition memory for faces was somewhat vulnerable to set size, but (c) recognition memory for pairs was most vulnerable, since the pattern of HR and FAR reversed with participants achieving very few Hits but very many False Alarms at high set sizes.

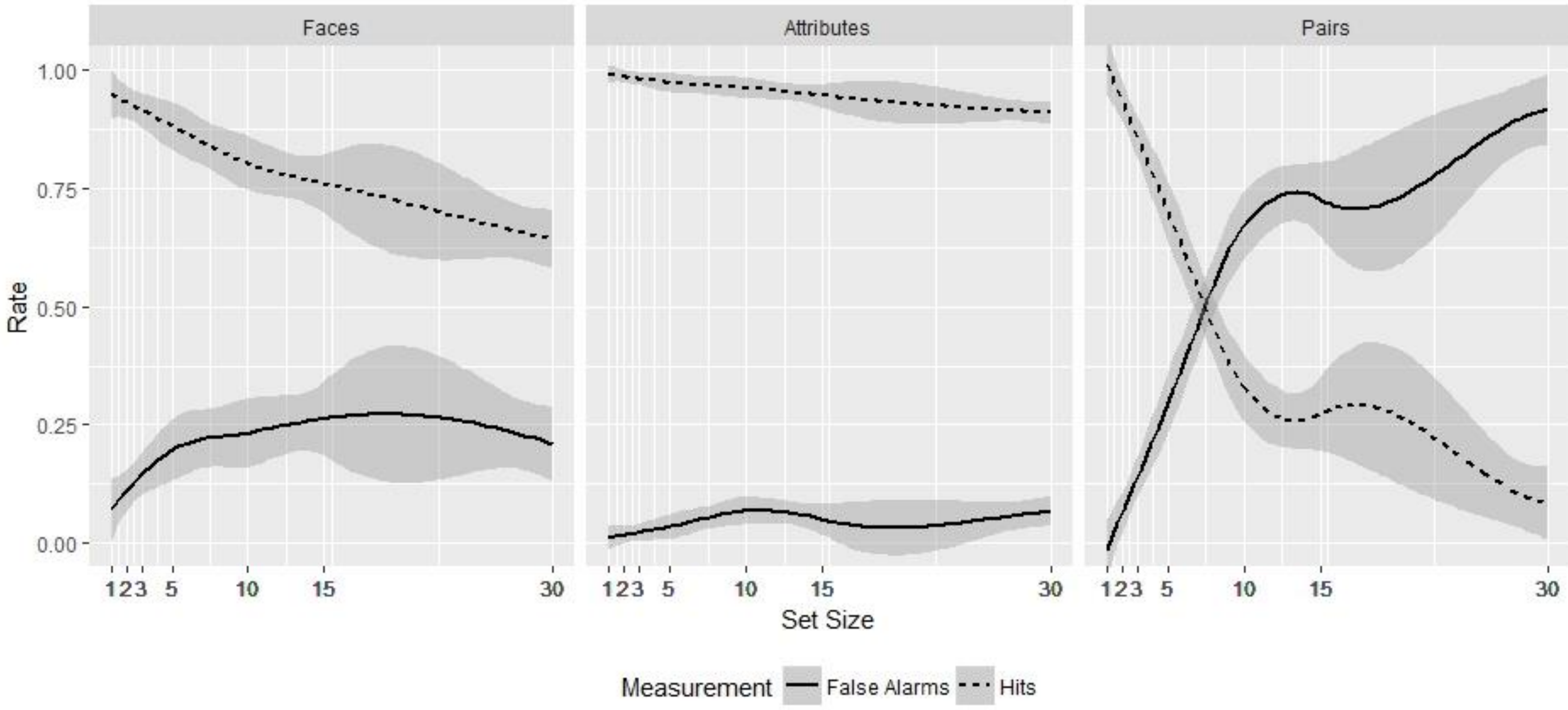

Figure 4.4. Mean Hit Rate and False Alarm Rate across Set Sizes (1,2,3,5,10, 15, and 30) for three tests: faces, attributes, and pairs. The shaded areas are 95% confidence intervals. Curves are LOESS non-parametric, span = 1.3. The solid black line represents false alarm rate, and the dashed line represents the hit rate.

HR, and FAR were analysed separately. Two mixed linear models were run with set size, test type, and the interaction between set size and test type as predictors (fixed effects) and performance for trials was nested within participants (random effects). All analyses were conducted in R (R Core Team, 2014), with the `lme4` package (Bates, Maechler, Bolker, & Walker, 2015). The `lme4` package produces various statistics about the model, including Akaike information criterion (AIC), Bayesian information criterion (BIC), loglikelihood, and deviance. The AIC and BIC statistics provide an estimate of how well the model fits the data (lower values mean a better fit), but the AIC is normally inflated from the number of data points (*n*) included in the model, and the BIC is conservative, because it penalises the model for each additional predictor. Instead, the parameters of the model are estimated by maximizing the likelihood function using the observed data values, and the negative logarithm of this statistic (-*LL*) is reported, and from this, the deviance is calculated (here deviance means the -*2Loglikelihood*). The difference between the deviance values of the two models is calculated when comparing two models to each other; these differences fall on a chi-square distribution. Even though the -*LL* is the more important statistic when evaluating the fit of the model, the AIC and BIC should support which model is a better fit and remain useful statistics. Models were further investigated using the `car` package (Fox & Weisberg, 2011). Both mixed linear model models will be discussed in turn below.

Mixed linear model for HR. First, a null model was constructed where HR was predicted by the intercept and the random intercepts for subject to account for the repeated measures design. HR was nested within subject. A second model was constructed that included the fixed effects set size and test type, and the random effect, subject. Finally, a third model was constructed with the same structure as the second model but also included a third fixed effect, the interaction term between set size and test type. Each model was compared to the subsequent model, and difference between two models' deviance score was tested with a chi-square test to determine which was the better fit for the data.

The model characteristics, and the outcome of the comparisons, are listed in Table 4.2. Model 2 significantly improved the fit of the data compared to the Null Model, χ^2 (8) = 215.99, p < .001, and Model 3 significantly improved the fit compared to Model 2, χ^2 (12) = 261.52, p < .001. For this reason, Model 3 was retained and further analysed. Two estimates of R^2 were calculated using the package `MuMIn` (Barton, 2018), which uses the methods outlined in Johnson (2014) and Nakagawa and Schielzeth (2013): $R^2_{marginal}$ estimated the proportion of variance explained by the fixed factors, and $R^2_{conditional}$ estimated the proportion of variance explained by the fixed and random factors Model 3 accounted for between 89.3% and 91.7% of the overall variance in the data ($R^2_{marginal}$ = 0.893, and $R^2_{conditional}$ = .917).

Table 4.2

Characteristics, and Chi-Square Tests of the Three Models Considered for Hits

Model Name	Model details	df	AIC	BIC	Log likelihood	Deviance	$\Delta\chi^2$	R_m^2	R_c^2
Null Model	Hits ~ 1 + (1 | Subject)	3	53.47	63.52	-23.74	47.47		0.000	0.073
Model 2	Hits ~ Set Size + Test + (1 | Subject)	11	-146.52	-109.70	84.260	-168.52	**215.99**	0.645	0.645
Model 3	Hits ~ Set Size *Test + (1 | Subject)	23	-384.03	-307.05	215.02	-430.03	**261.52**	0.893	0.917

Note. This table shows the characteristics of the three, different mixed linear models. The column 'Model details' lists the syntax for each model. The tilde symbol (~) denotes prediction, the number one '1' denotes the intercept and '1|' indicates random effects (i.e., intercepts) grouped by the subsequent term. Therefore, for the null model, the outcome variable Hits is predicted by the intercept, and by random effects for each subject. The asterisk in the model details for Model 3 indicates that Hits is predicted by both variables and their interaction term.
Embolden text denotes significance ($p < .05$). The df for the chi-square statistic is the difference in the df of the two models.
R_m^2 and R_c^2 denote R-squared for fixed factors, and fixed and random factors respectively. R_m^2 is zero for the Null Model, because this model contained no fixed factors.

In Model 3, HR was predicted by set size, test type, and the interaction term between set size and test type. A closer examination of Model 3 showed that there was a main effect for both set size, χ^2 (6) = 447.62, p < .001, and test, χ^2 (2) = 792.06, p <.001, and a significant interaction between set size and test type, χ^2 (12) = 627.32, p < .001. The model coefficients are listed in Table 4.3. Each main effect, and the interaction effect, will be discussed in turn.

Table 4.3

Coefficients for Fixed Effects and Random Effects for Model 3 Predicting Hits

Fixed Effects	B	CI lower	CI upper	β	CI lower	CI upper	p
Intercept (Set Size 1, Attributes Test)	0.99	0.93	1.05	-	-	-	**<.001**
SetSize2	0.00	-0.07	0.08	0.00	-0.10	0.10	.932
SetSize3	-0.01	-0.09	0.07	-0.01	-0.11	0.09	.804
SetSize5	-0.02	-0.09	0.06	-0.02	-0.12	0.08	.677
SetSize10	-0.03	-0.10	0.05	-0.03	-0.13	0.07	.503
SetSize15	-0.04	-0.12	0.03	-0.06	-0.16	0.04	.279
SetSize30	-0.08	-0.16	-0.00	-0.10	-0.20	-0.00	**.046**
Faces Test	-0.05	-0.12	0.02	-0.09	-0.21	0.03	.130
Pairs Test	0.00	-0.07	0.07	0.01	-0.11	0.13	.923
Set Size2: Faces Test	0.01	-0.09	0.10	0.01	-0.07	0.08	.889
SetSize3: Faces Test	0.00	-0.09	0.10	0.00	-0.07	0.08	.949
SetSize5: Faces Test	-0.05	-0.15	0.05	-0.04	-0.12	0.04	.315
SetSize10: Faces Test	-0.10	-0.20	-0.01	-0.08	-0.16	0.00	**.039**
SetSize15: Faces Test	-0.13	-0.23	-0.04	-0.10	-0.18	-0.03	**.008**
SetSize30: Faces Test	-0.21	-0.31	-0.12	-0.17	-0.24	-0.09	**<.001**
SetSize2: Pairs Test	-0.03	-0.13	0.06	-0.03	-0.10	0.05	.500
SetSize3: Pairs Test	-0.12	-0.21	-0.02	-0.09	-0.17	-0.02	**.020**
SetSize5: Pairs Test	-0.29	-0.38	-0.19	-0.22	-0.30	-0.15	**<.001**
SetSize10: Pairs Test	-0.64	-0.74	-0.55	-0.50	-0.58	-0.43	**<.001**
SetSize15: Pairs Test	-0.68	-0.77	-0.58	-0.53	-0.61	-0.45	**<.001**
SetSize30: Pairs Test	-0.83	-0.93	-0.73	-0.65	-0.73	-0.57	**<.001**
Random Effects							
Participants							
Number of observations	210						
Number of participants	70						
ICC	0.224						
Variance	0.006						

Note. CI stands for confidence interval. The reference group for the fixed effect, Set Size, is Set Size 1. The reference group for the fixed effect, Test Type, is Attribute. The reference group for the interaction between Set Size and Test Type, is Set Size 1: Attribute test. Embolden text indicates significant p-values at α =.05.

There was a significant main effect of set size on HR, χ^2 (6) = 447.62, p < .001. This main effect was best fitted with a linear, t(70) = -20.581, p <.001, quadratic, t(70) = -3.649, p<.001, or cubic trend, t(70) = 2.255, p =.027. Inspection of Figure 4.5 suggests that there may be two inflection points, one at Set Size 5 and another at Set Size 15 (supporting the cubic trend) with the HR performance approaching an asymptote of roughly .50 (i.e., chance). Further contrasts revealed that there was no significant difference in HR between Set Sizes 1 and 2, or between Set Sizes 2 and 3, all ps > 0.05. HR was significantly higher for Set Size 3 than Set Size 5, t(70) = -2.935, p = 0.025, $\overline{HR}_{Set\ Size\ 3}$ = 0.926, 95% CI [0.89,0.97], SE = 0.02. versus $\overline{HR}_{Set\ Size\ 5}$ = 0.845, 95% CI [0.81,0.88], SE = 0.02. Performance continued to decline: Participants performed significantly worse for Set Size 10 than Set Size 5, t(70) = -5.313, p < .001, $\overline{HR}_{Set\ Size\ 5}$ = 0.845, 95% CI [0.81,0.88], SE = 0.02 versus $\overline{HR}_{Set\ Size\ 10}$ = 0.698, 95% CI [0.659,0.737], SE = 0.02. There was no significant difference in HR between Set Sizes 10 and 15, p > .05, but participants performed significantly worse for Set Size 30 than Set Size 15, t(70) = -4.145, p < .001, $\overline{HR}_{Set\ Size\ 15}$ = 0.660, 95% CI [0.621,0.699], SE = 0.02 versus $\overline{HR}_{Set\ Size\ 30}$ = 0.546, 95% CI [0.507,0.585], SE = 0.02. These results are shown in Figure 4.5.

The main effect of test type was significant, χ^2 (2) = 792.06, p <.001. HR was significantly worse for faces than for attributes, t(140) = 9.316, p <.001, $\overline{HR}_{Faces}$= 0.84, 95% CI [0.82,0.86], SE = .01 versus $\overline{HR}_{Attributes}$= 0.97, 95% CI [0.94,0.99], SE = .01, and HR was significantly worse for pairs than for faces, t(140) = 18.341, p < .001 $\overline{HR}_{Pairs}$= = 0.60, 95% CI [0.58,0.62], SE = .01 versus $\overline{HR}_{Faces}$= = 0.84, 95% CI [0.82,0.86], SE = .01. These results are shown in Figure 4.5.

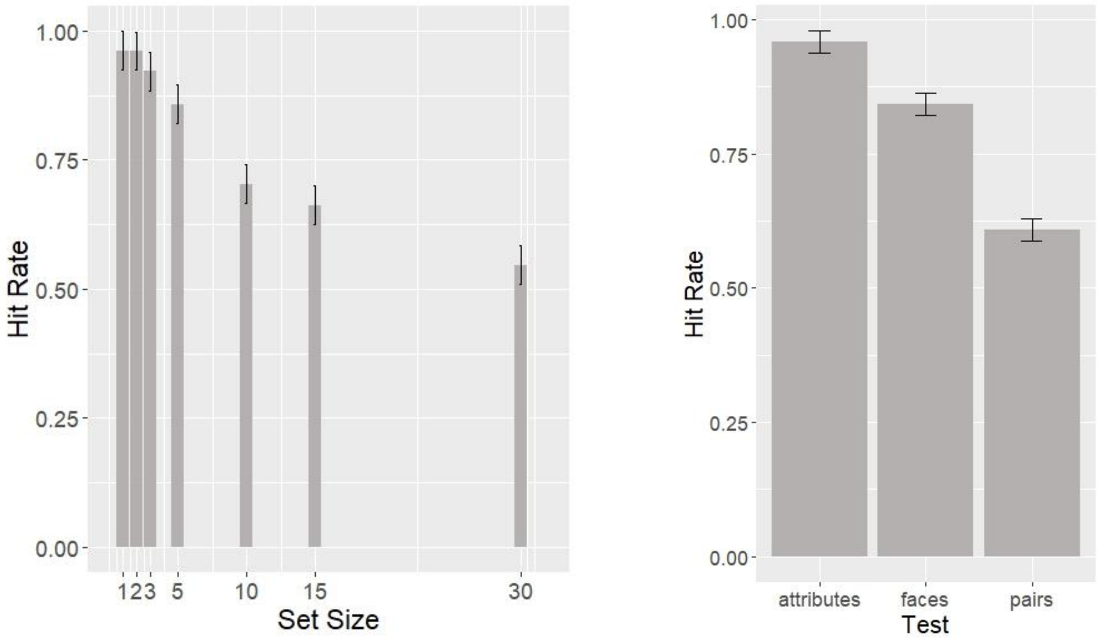

Figure 4.5 Hit Rate across set size and test type. The image on the left shows HR as a function of set size, and the image on the right shows HR as a function of test type. The errors bars are 95% confidence intervals.

The interaction between test type and set size was significant, χ^2 (12) = 627.32, $p <$ .001 (see Figure 4.6). There was no significant difference in HR between the three test types at Set Size 1, Set Size 2, or Set Size 3, $p > 0.05$. At Set Size 5, HR was significantly worse for pairs than for attributes, and significantly worse for pairs than for faces; but there was no significant difference between HR for faces and attributes. From Set Size 10 onwards, HR differed significantly between all three test types. This is evident in Figure 4.6, and post-hoc test results are listed in Table 4.4. All p-values were adjusted for multiple pairwise comparisons.

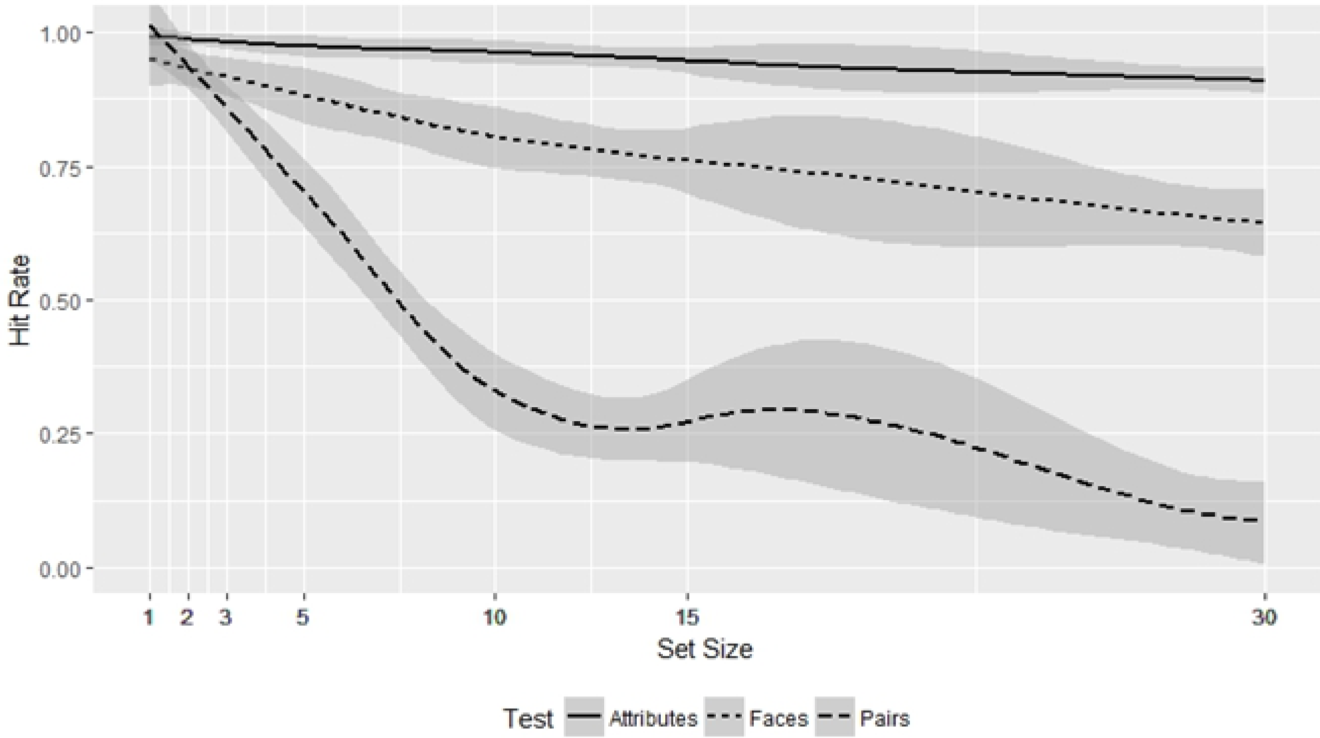

Figure 4.6. Mean HR across set sizes (1,2,3,5,10, 15, and 30) for three tests (faces, attributes, and pairs). The shaded areas are 95% confidence intervals. Curves are LOESS non-parametric, span = 1.3. The solid black line represents the results for the attributes test, the dashed line shows the results for the faces test, and the long-dash line represents the results for the pairs test.

Table 4.4
Pairwise Comparisons between HR of Each Test within Each Set Size.

Set Size	Test	t	df	p
1	Attributes vs Faces	1.52	140.0	.995
	Attributes vs Pairs	-0.10	140.0	1.000
	Pairs vs Faces	-1.62	140.0	.989
2	Attributes vs Faces	1.33	140.0	.999
	Attributes vs Pairs	0.86	140.0	1.000
	Pairs vs Faces	-0.47	140.0	1.000
3	Attributes vs Faces	1.43	140.0	.998
	Attributes vs Pairs	3.24	140.0	.151
	Pairs vs Faces	1.81	140.0	.965
5	Attributes vs Faces	2.95	140.0	.291
	Attributes vs Pairs	8.09	140.0	**<.0001**
	Pairs vs Faces	5.14	140.0	**.0002**
10	Attributes vs Faces	4.47	140.0	**.0028**
	Attributes vs Pairs	18.27	140.0	**<.0001**
	Pairs vs Faces	13.79	140.0	**<.0001**
15	Attributes vs Faces	5.33	140.0	**<.0001**
	Attributes vs Pairs	19.22	140.0	**<.0001**
	Pairs vs Faces	13.89	140.0	**<.0001**
30	Attributes vs Faces	7.61	140.0	**<.0001**
	Attributes vs Pairs	23.60	140.0	**<.0001**
	Pairs vs Faces	15.99	140.0	**<.0001**

Notes. P-values are adjusted for pairwise error with Tukey's procedure. Embolden terms indicate significance at $\alpha = .05$.

Mixed linear model for FAR. A linear mixed model was built to predict FAR. Like the model for HR, random intercepts were estimated for participants; the random intercepts captured the repeated-measures design of the experiment. FAR was nested within the random effect, subject. Three models were built: First, a null model where FAR was predicted by the intercept and random effect, subject; a second model where FAR was predicted by the fixed effects, test type and set size, and the random effect, subject; and a third model where FAR was predicted by the fixed effects, test type, sets, an interaction term between test type and set size, and random effect, subject. Each model was compared with the subsequent model. The results and model characteristics are shown in Table 4.5.

Table 4.5.

Characteristics, and Chi-Square Tests of the Three Models that Predict False Alarm Rate

Model Name	Model details	df	AIC	BIC	Log likelihood	Deviance	$\Delta\chi^2$	R_m^2	R_c^2
Null Model	FAR ~ 1 + (1 | Subject)	3	51.774	61.785	-22.872	45.744		0.000	0.000
Model 2	FAR ~ Set size + Test + (1 | Subject)	11	-111.287	-74.469	66.643	-133.287	**179.03**	0.575	0.575
Model 3	FAR ~ Set size *Test + (1 | Subject)	23	-330.64	-253.655	188.319	-376.64	**243.35**	0.861	0.890

Notes. This table shows the characteristics of the three different multilevel models. The column ‘Model details’ lists the syntax for each model. The tilde symbol (~) denotes prediction, the number one ‘1’ denotes the intercept and ‘1|’ indicates random effects (i.e., intercepts) grouped by the subsequent term. Therefore, for the null model, the outcome variable False Alarms is predicted by the intercept, and by random effects for each subject. The asterisk in the model details for Model 3 indicates that Hits is predicted by both variables, and their interaction term.

FAR denotes False Alarm Rate.

Embolden text denotes significance (p < .05). The df for the chi-square statistic is the difference in the df of the two models.

R_m^2 and R_c^2 denote R-squared and R-squared for fixed factors, and fixed and random factors respectively. R_m^2 is zero for the Null Model, because there are no fixed factors in this model.

Model 2 was a significantly better fit than the null model, χ^2 (8) = 179.03, p < .001, and Model 3 was a significantly better fit than Model 2, χ^2 (12) = 243.35, p < .001. Thus, further analyses were conducted on Model 3. Overall, Model 3 accounted for between 86.1% and 89% of the variance ($R^2_{marginal}$ = 0.861, and $R^2_{conditional}$ = .890). The model coefficients for Model 3 are reported in Table 4.6.

Table 4.6.

Coefficients for Fixed Effects and Random Effects for Model 3 that Predicts False Alarm Rate

Fixed Effects	B	CI lower	CI upper	β	CI lower	CI upper	*p*
Intercept (Set Size 1, Attributes Test)	0.01	-0.05	0.08	-	-	-	.677
SetSize2	0.00	-0.09	0.09	0.00	-0.11	0.11	.998
SetSize3	0.01	-0.08	0.10	0.02	-0.10	0.13	.770
SetSize5	0.02	-0.07	0.11	0.03	-0.09	0.14	.656
SetSize10	0.06	-0.03	0.14	0.07	-0.04	0.19	.210
SetSize15	0.04	-0.05	0.12	0.05	-0.07	0.16	.416
SetSize30	0.05	-0.04	0.14	0.07	-0.05	0.18	.239
Faces Test	0.05	-0.03	0.13	0.09	-0.04	0.23	.184
Pairs Test	-0.01	-0.09	0.07	-0.01	-0.15	0.13	.867
Set Size2: Faces Test	0.05	-0.06	0.16	0.04	-0.05	0.13	.378
SetSize3: Faces Test	0.07	-0.04	0.18	0.05	-0.03	0.14	.241
SetSize5: Faces Test	0.11	0.00	0.22	0.08	0.00	0.17	.062
SetSize10: Faces Test	0.11	0.00	0.22	0.08	0.00	0.17	.062
SetSize15: Faces Test	0.16	0.05	0.27	0.13	0.04	0.21	**.005**
SetSize30: Faces Test	0.09	-0.02	0.20	0.07	-0.02	0.16	.115
SetSize2: Pairs Test	0.03	-0.08	0.14	0.02	-0.06	0.11	.595
SetSize3: Pairs Test	0.11	0.00	0.22	0.09	0.00	0.18	**.047**
SetSize5: Pairs Test	0.28	0.17	0.39	0.22	0.14	0.31	**<.001**
SetSize10: Pairs Test	0.61	0.50	0.72	0.48	0.40	0.57	**<.001**
SetSize15: Pairs Test	0.68	0.57	0.79	0.54	0.45	0.63	**<.001**
SetSize30: Pairs Test	0.86	0.75	0.97	0.68	0.59	0.76	**<.001**
Random Effects							
Participants							
Number of observations	210						
Number of participants	70						
ICC	0.224						
Variance	0.008						

Note. ICC denotes interclass correlation.
P values are adjusted for pairwise error with Tukey's procedure.
Embolden terms indicate significance at α = .05.

There were significant main effects for both set size, χ^2 (6) = 284.30, $p < .001$, and test type, χ^2 (2) = 583.84, $p < .001$. The results of the polynomial contrast showed that FAR across set size was best fit with a linear, $t(70) = 16.501$, $p < .001$, or cubic trend, $t(70) = -2.716$, $p = .0083$. Inspection of Figure 4.7 suggests one inflection point at Set Size 10, and perhaps a second inflection point at Set Size 15. Consecutive contrasts showed that the only significant difference between pairwise comparisons of set size was between Set Sizes 5 and 10, $t(70) = 4.735$, $p < .001$. These results are shown in Figure 4.7.

The main effect of test type on FAR was further analysed. Consecutive contrasts showed that the FAR differed significantly between attributes and faces, $t(140) = -9.002$, $p < .0001$, attributes and pairs, $t(140) = -23.920$, $p < .0001$, and faces and pairs, $t(140) = -14.918$, $p < .0001$. These results are shown in Figure 4.7.

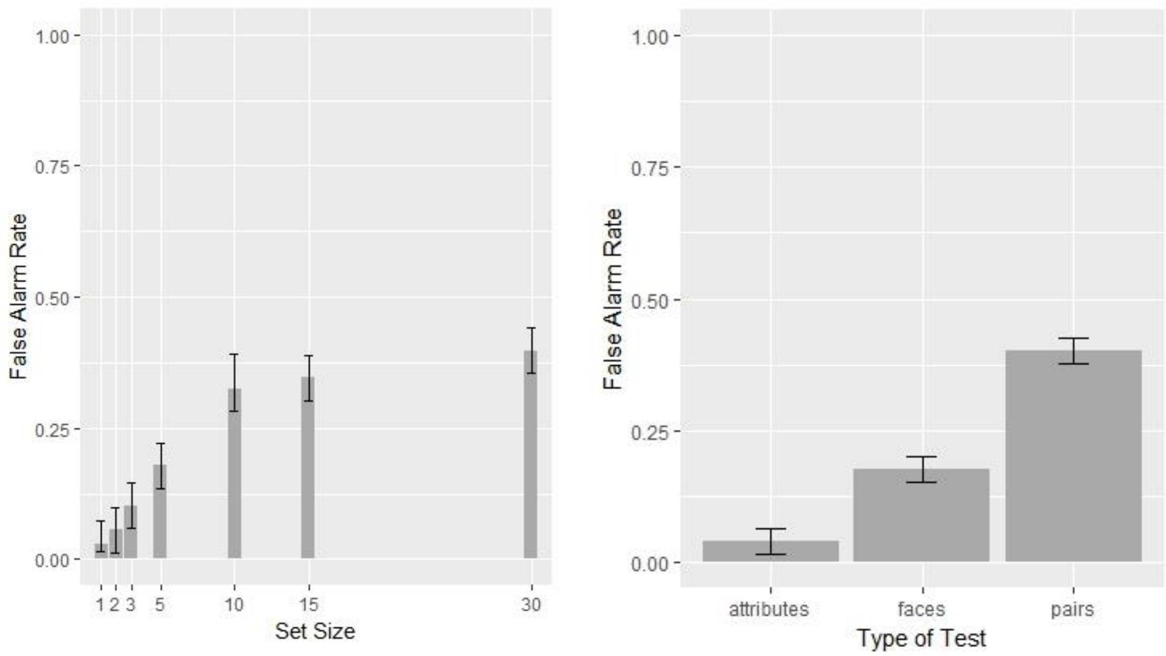

Figure 4.7. Main effects of set size, and test type on FAR. The image on the left shows the main effect of set size on FAR. The image on the right shows the main effect of test type on FAR. The errors bars are 95% confidence intervals.

The interaction term between set size and test type was significant, χ^2 (12) = 547.51, $p < .001$. The contrasts are shown in Table 4.7, and the results are displayed in Figure 4.8. The contrasts suggested no significant difference between FAR of the three tests at Set Sizes 1, 2, and 3, all *p*s > .05. At Set Size 5, the differences in FAR were significant between attributes and faces, and between attributes and pairs; that is, the FAR was significantly higher for both Faces, and Pairs than for Attributes. The difference between FAR for pairs and faces was not significant. At Set Size 10 and 15, FAR for all three tests differed significantly from each other, and at Set Size 30, FAR for pairs remained significantly different from FAR for faces and attributes, but the difference in FAR for faces and attributes was no longer statistically significant.

Table 4.7

Pairwise Comparisons between FAR of Each Test within Each Set Size.

Set Size	Test	t	df	p
1	Attributes vs Faces	-1.33	140	.999
	Attributes vs Pairs	0.17	140	1.000
	Pairs vs Faces	1.50	140	.996
2	Attributes vs Faces	-2.59	140	.545
	Attributes vs Pairs	-0.59	140	1.000
	Pairs vs Faces	2.00	140	.911
3	Attributes vs Faces	-3.00	140	.263
	Attributes vs Pairs	-2.67	140	.481
	Pairs vs Faces	0.33	140	1.00
5	Attributes vs Faces	-4.00	140	**.016**
	Attributes vs Pairs	-6.91	140	**<.001**
	Pairs vs Faces	-2.91	140	.314
10	Attributes vs Faces	-4.00	140	**.016**
	Attributes vs Pairs	-15.15	140	**<.001**
	Pairs vs Faces	-11.16	140	**<.001**
15	Attributes vs Faces	-5.33	140	**<.001**
	Attributes vs Pairs	-16.90	140	**<.001**
	Pairs vs Faces	-11.57	140	**<.001**
30	Attributes vs Faces	-3.54	140	.060
	Attributes vs Pairs	-21.24	140	**<.001**
	Pairs vs Faces	-17.66	140	**<.001**

Notes. P values are adjusted for pairwise error with Tukey's procedure.
Embolden terms indicate significance at $\alpha = .05$.

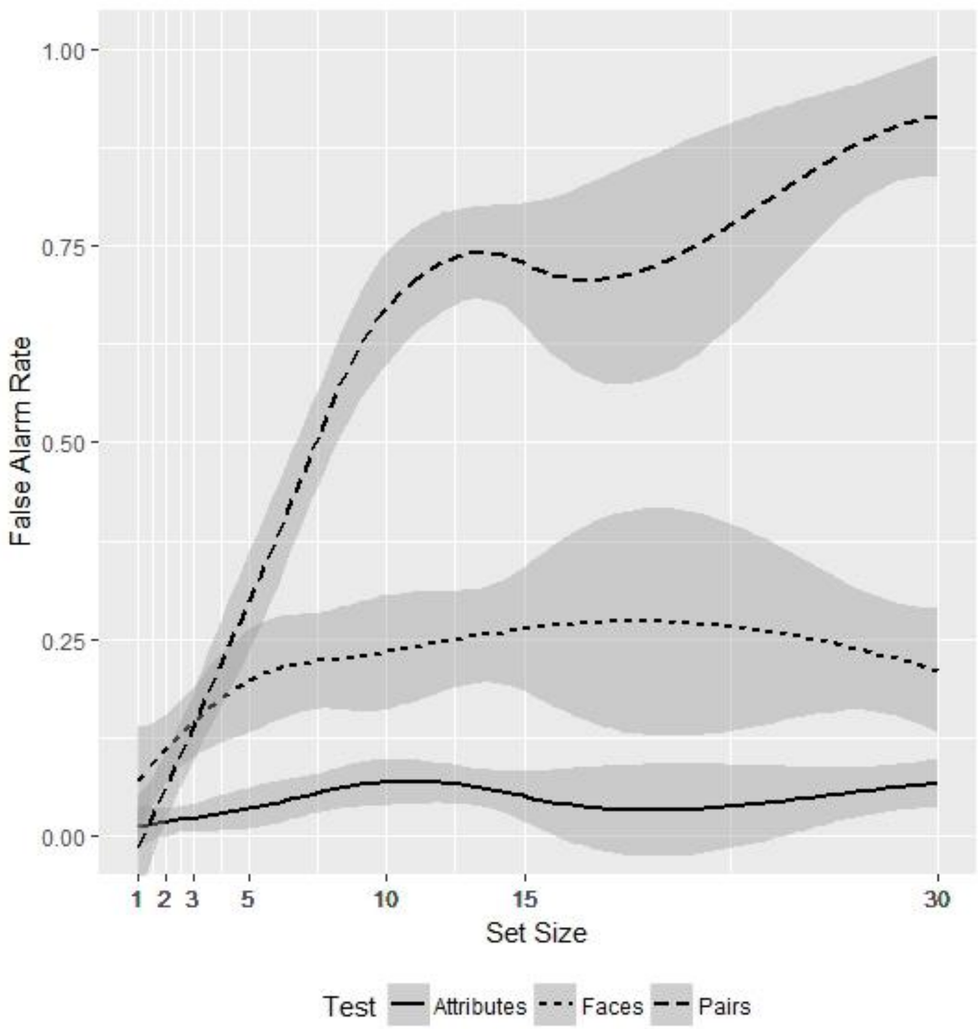

Figure 4.8. Mean FAR across set sizes (1,2,3,5,10, 15, and 30) for three tests, faces, attributes, and pairs. The shaded areas are 95% confidence intervals. Curves are LOESS non-parametric, span = 1.3. The solid black line represents the results for the attributes test, the dashed line shows the results for the faces test, and the long-dash line represents the results for the pairs test.

Confidence

Participants rated their confidence in their decision for each item in the faces, attributes, and pairs test, on a scale from 0% to 100%. The average confidence scores for each of the three tests across each set size are reported in Table 4.8. Inspection of Table 4.8 shows that the overall average confidence is highest for attributes, followed by faces, and then pairs.

Table 4.8.

Average Confidence for the Three Tests Across Set Size

Set Size	Faces	Attributes	Pairs
1	91.29 (0.75)	98.98 (0.39)	96.01 (0.84)
2	91.60 (0.68)	99.25 (0.26)	96.84 (0.65)
3	88.49 (1.02)	97.71 (0.49)	92.36 (0.78)
5	87.59 (0.84)	97.41 (0.41)	86.95 (1.23)
10	77.64 (0.79)	93.72 (0.50)	63.50 (1.40)
15	80.20 (0.88)	93.99 (0.62)	52.15 (1.58)
30	74.40 (1.03)	92.01 (0.72)	31.94 (1.54)
Overall Average	84.46 (0.20)	96.11 (0.33)	74.25 (0.69)

Notes. Values in parentheses indicate standard errors.

The repeated measures correlation (Bakdash & Marusich, 2017) showed a moderately strong and positive relationship between confidence and accuracy, r = 0.46, 95% CI [0.44, 0.48], $p < .001$.

Mixed linear model for confidence. Three models to predict confidence were constructed. Confidence was nested within subject. These three models followed the same structure as those for FAR and HR: The null model included only the intercept and the random effect, subject; the second model included the fixed effects set size and test type and the random effect, subject; and the third model included the fixed effects set size and test type, and an interaction term between set size and test type, including the random effect, subject. The model structure and model statistics are reported in Table 4.9. Model 2 was a better fit than Model 1,χ^2 (8) = 153.04, $p < .001$, and Model 3 was a better fit than Model 2, χ^2 (12) = 173.21, $p < .001$. The fixed factors of Model 3 accounted for 76.3% of the variance, and this increased to 85.9% when the random effects were included, R_m^2 = .763 and R_c^2 = .859. Further analyses were conducted on Model 3.

Table 4.9

Characteristics, and Chi-Square Tests of the Three Models Considered for Confidence

Model Name	Model details	df	AIC	BIC	Log likelihood	Deviance	$\Delta\chi^2$	R_m^2	R_c^2
Null Model	Confidence ~ 1 + (1 | Subject)	3	1835.3	1845.4	-914.67	1829.3		0.000	0.189
Model 2	Confidence ~ Set size + Test + (1 | Subject)	11	1698.3	1735.1	-838.15	1676.3	**153.04**	0.534	0.534
Model 3	Confidence ~ Set size *Test + (1 | Subject)	23	1549.1	1626.1	-751.54	1503.1	**173.21**	0.763	0.859

Notes. This table shows the characteristics of the three different multilevel models. The column ‘Model details’ lists the syntax for each model. The tilde symbol (~) denotes prediction, the number one ‘1’ denotes the intercept and ‘1|’ indicates random effects (i.e., intercepts) grouped by the subsequent term. Therefore, for the null model, the outcome variable confidence is predicted by the intercept, and by random effects for each subject. The asterisk in the model details for Model 3 indicates that Confidence is predicted by both variables, and their interaction term.
Embolden text denotes significance (p < .05). The df for the chi-square statistic is the difference in the df of the two models. R_m^2 and R_c^2 denote R-squared and R-squared for fixed factors, and fixed and random factors respectively. R_m^2 is zero for the Null Model, because there are no fixed factors in this model.

The coefficients of Model 3 are reported in Table 4.10. There was a significant main effect for set size, χ^2 (6) = 154.06, p < .001, and test type, χ^2 (2) = 323.24, p < .001, and a significant main interaction between set size and test type, χ^2 (12) = 341.32, p < .001

Table 4.10.

Model Coefficients for Model 3 Predicting Confidence

Fixed Effects	B	CI lower	CI upper	β	CI lower	CI upper	p
Intercept (Set Size 1, Attributes Test)	98.69	92.90	100.00	-	-	-	<.001
SetSize2	0.56	-7.63	8.75	0.01	-0.14	0.16	.894
SetSize3	-0.99	-9.17	7.20	-0.02	-0.17	0.13	.814
SetSize5	-1.28	-9.47	6.91	-0.02	-0.17	0.13	.759
SetSize10	-4.97	-13.16	3.21	-0.09	-0.24	0.06	.236
SetSize15	-4.70	-12.89	3.49	-0.09	-0.24	0.06	.262
SetSize30	-6.68	-14.87	1.51	-0.12	-0.27	0.03	.112
Faces Test	-7.40	-13.71	-1.10	-0.18	-0.34	-0.03	**.023**
Pairs Test	-2.68	-8.99	3.63	-0.07	-0.22	0.09	.407
Set Size2: Faces Test	-0.24	-9.17	8.68	0.00	-0.10	0.10	.957
SetSize3: Faces Test	-1.80	-10.72	7.13	-0.02	-0.12	0.08	.694
SetSize5: Faces Test	-2.42	-11.34	6.51	-0.03	-0.13	0.07	.597
SetSize10: Faces Test	-8.68	-17.60	0.25	-0.10	-0.20	0.00	.059
SetSize15: Faces Test	-6.39	-15.31	2.53	-0.07	-0.17	0.03	.163
SetSize30: Faces Test	-10.21	-19.13	-1.29	-0.11	-0.21	-0.01	**.027**
SetSize2: Pairs Test	0.27	-8.65	9.19	0.00	-0.10	0.10	.953
SetSize3: Pairs Test	-2.67	-11.59	6.26	-0.03	-0.13	0.07	.559
SetSize5: Pairs Test	-7.78	-16.70	1.14	-0.09	-0.1	0.01	.090
SetSize10: Pairs Test	-27.54	-36.47	-18.62	-0.31	-0.41	-0.21	**<.001**
SetSize15: Pairs Test	-39.17	-48.09	-30.25	-0.44	-0.53	-0.34	**<.001**
SetSize30: Pairs Test	-57.39	-66.31	-48.47	-0.64	-0.74	-0.54	**<.001**
Random Effects							
Participants							
Number of observations	210						
Number of participants	70						
ICC	0.224						
Variance	0.008						

Note. ICC denotes interclass correlation.
P values are adjusted for pairwise error with Tukey’s procedure.
Embolden terms indicate significance at α = .05.

For the main effect of set size, the results of the polynomial contrast showed that confidence across set size was best fit with a quadratic trend, $t(70) = -3.269$, $p = .0017$. Consecutive contrasts showed that the only significant difference in confidence was between Set Sizes 5 and 10, $t(70) = -3.808$, $p = .002$, and Set Sizes 15 and 30, $t(70) = -2.873$, $p = .027$ (see Figure 4.9).

For the main effect of test type, consecutive contrasts showed that confidence differed significantly between attributes and faces, $t(140) = 9.576$, $p < .001$, and between faces and pairs, $t(140) = 8.390$, $p < .001$. Confidence was highest for the attribute test, followed by the faces and pairs test respectively (see Figure 4.9).

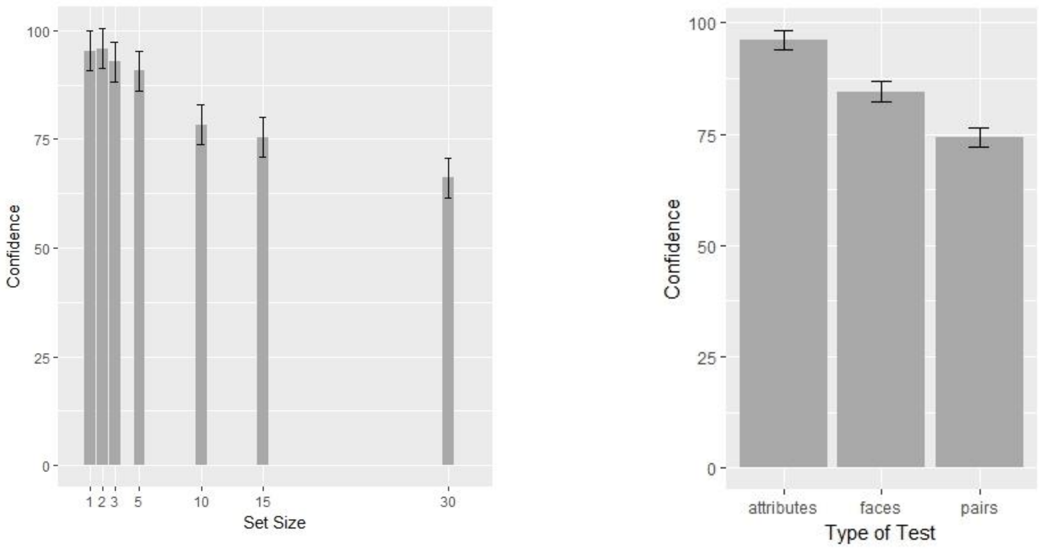

Figure 4.9. Main effects of set size and test type on confidence. The image on the left shows the main effect of set size on confidence. The image on the right shows the main effect of test type on confidence. The errors bars are 95% confidence intervals.

For the interaction between set size and test type, pairwise comparisons showed no significant difference between confidence levels of each test type within Set Size 1 up to Set Size 5. However, from Set Size 10 onwards, confidence differed between each test type consistently, with significantly lower confidence for pairs decisions, than for faces or attribute decisions. These results and pairwise comparisons are in Figure 4.10, and Table 4.11 respectively.

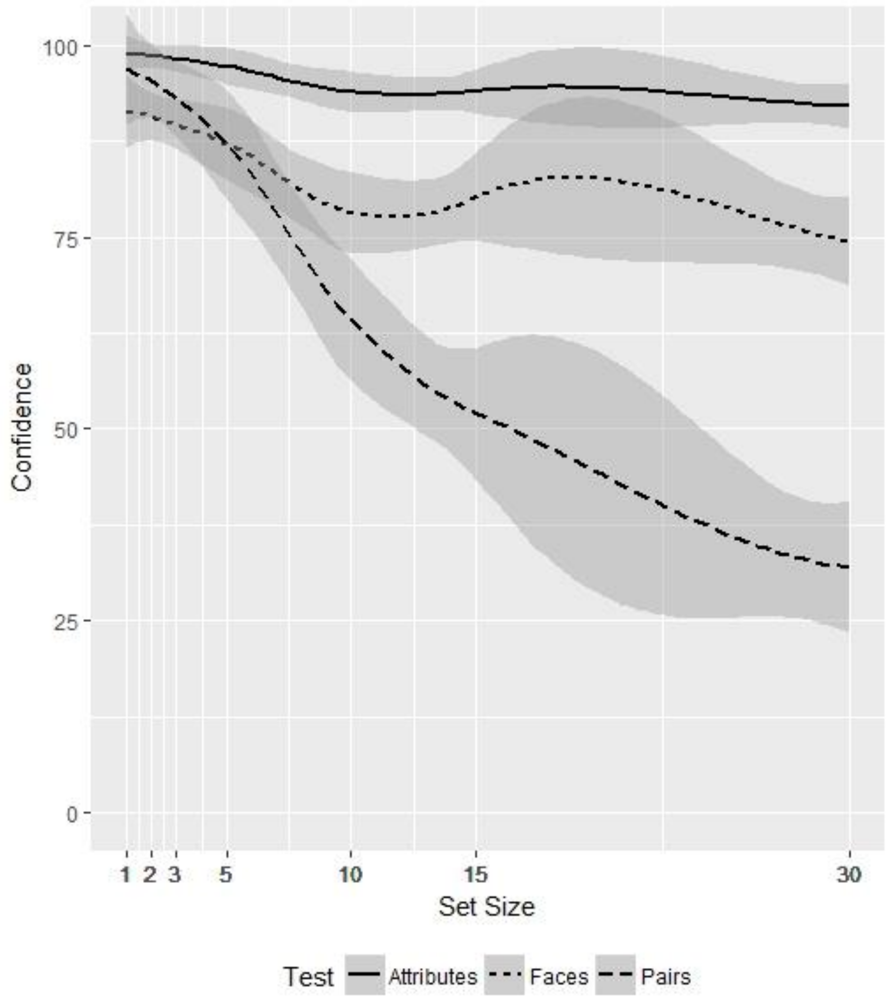

Figure 4.10. Mean confidence across Set Sizes (1,2,3,5,10, 15, and 30) for three tests, faces, attributes, and pairs. The shaded areas are 95% confidence intervals. Curves are LOESS non-parametric, span = 1.3. The solid black line represents the results for the attributes test, the dashed line shows the results for the faces test, and the long-dash line represents the results for the pairs test.

Table 4.11

Pairwise Comparisons Between Mean Confidence of Each Test Within Each Set Size.

Set Size	Test	t-test	df	p
1	Attributes vs Faces	2.30	140	.753
	Attributes vs Pairs	0.83	140	1.000
	Pairs vs Faces	-1.47	140	.997
2	Attributes vs Faces	2.38	140	.701
	Attributes vs Pairs	0.748	140	1.000
	Pairs vs Faces	-1.63	140	.988
3	Attributes vs Faces	2.858	140	.349
	Attributes vs Pairs	1.66	140	.986
	Pairs vs Faces	-1.20	140	1.000
5	Attributes vs Faces	3.05	140	.236
	Attributes vs Pairs	3.25	140	.148
	Pairs vs Faces	0.199	140	1.000
10	Attributes vs Faces	5.00	140	**<.001**
	Attributes vs Pairs	9.39	140	**<.001**
	Pairs vs Faces	4.39	140	**.004**
15	Attributes vs Faces	4.29	140	**.006**
	Attributes vs Pairs	13.00	140	**<.001**
	Pairs vs Faces	8.71	140	**<.001**
30	Attributes vs Faces	5.47	140	**<.001**
	Attributes vs Pairs	18.66	140	**<.001**
	Pairs vs Faces	13.19	140	**<.001**

Note. P values are adjusted for pairwise error with Tukey's procedure.
Embolden terms indicate significance at $\alpha = .05$.

Discussion

The primary conclusion from this experiment is that set size does impact recognition performance, but not consistently across all three tests, and not immediately at the lowest set size. The results from this experiment showed that both HR and FAR were affected by set size. Participants performed worse (i.e., lower HR, and higher FAR) as set size increased. This decrease in performance was not uniform across all three types of tests: Memory for attributes was robust and largely unaffected by set size, memory for faces was more vulnerable and showed some decline in performance, and memory for pairs was most sensitive to the negative impact of increased set sizes.

The marked decrease in recognition performance for pairs is striking, considering that memory for faces and attributes were somewhat robust. In fact, at Set Size 30, HR for faces and attributes was roughly 64% and 91% respectively, whereas HR for pairs was 2%. The difference in recognition performance across the three tests suggest that associative memory is independent of its individual components.

Furthermore, there was no difference in recognition performance between the three tests at lower Set Sizes (< 5). At Set Size 5, however, significant differences in recognition were evident: Compared to attributes, the FAR was significantly higher for faces and for pairs, and compared to pairs, HR was significantly higher for attributes and faces. From Set Size 10 onwards, HR and FAR differed significantly among all three tests. These results suggest that human memory is unaffected by load at encoding at lower set sizes, but associative memory, and its individual components that comprise this memory, are more vulnerable to higher set sizes, albeit to varying degrees.

The negative effect of set size extended to self-reported measures of confidence. This effect was also not uniform across the three tests, and was not evident at lower set sizes. Overall, participants reported lower confidence following the pairs test, than the faces and attributes tests. There were no significant differences between the three tests at lower set sizes (<10), but the differences in confidence for the three tests was significant from Set Size 10 onwards.

Unlike the studies reported by Podd (1990), Metzger (2002), and Lamont et al. (2005), this experiment included trials with lower set sizes and tested recognition performance for all encoding stimuli. Like the results from Podd (1990) and Metzger (2002), the results of the current study suggest that set size had a detrimental effect on recognition performance. The results from the current experiment are in line with the results reported by Metzger (2002): Set size had a significant effect on both HR and FAR (see Table 4.12). What remains unclear from the results of the current experiment is how recognition performance will continue to change at even higher set sizes, like those used by Podd (1990), and Lamont et al. (2005). Will the performance continue to decrease? Will this decrease be linear? Recognition performance for pairs had already reached 2% HR and 98% FAR at Set Size 30, and will likely remain unchanged, but it remains unknown how recognition performance for faces and attributes will change.

Table 4.12

A comparison of the Face Recognition Results of Face Recognition Experiment 1, Podd (1990), Metzger, 2002, and Lamont et al. (2005)

	Face Recognition Experiment 1		Podd (1990)		Metzger (2002)		Lamont et al. (2005)	
SS	HR	FAR	HR	FAR	HR	FAR	HR	FAR
1	.94	.07	-	-	-	-	-	-
2	.95	.12	-	-	-	-	-	-
3	.93	.15	-	-	-	-	-	-
5	.87	.19	-	-	-	-	-	-
10	.81	.23	-	-	.79	.17	-	-
15	.76	.26	-	-	-	-	-	-
20	-	-	.74	.20	.72	.20	.78	.19
30	.64	.21			.67	.25	-	-
35	-	-	.68	.23	-	-	-	-
40	-	-			-	-	.78	.22
50	-	-	.65	.24	-	-	-	-

Note. SS indicates Set Size. HR and FAR denotes Hit Rate and False Alarm Rate respectively.
Only the HR and FAR for the faces test are reported in this table since Podd (1990), Metzger (2002), and Lamont et al. (2005) only tested recognition of faces.

Unlike previous research that found superior recognition performance for visual memory compared to verbal memory (Shepard, 1967), the results from Face Recognition Experiment 1 demonstrated that verbal memory was superior to visual memory. The impaired performance for visual memory is most likely due to face images being a very specific type of visual memory - one that is more fallible to the effects of Set Size. Additionally, recognition memory for the attributes may be easier because there is less variation within the item between encoding and recognition, but also because the attributes were presented visually. Recognition performance for verbal items might change due to modality at presentation, for example, a sentence presented visually or aurally (Crottaz-Herbette, Anagnoson, & Menon, 2004; Gunter, Furnham, & Gietson, 1984).

The results for the pairs test were surprising. Despite being given only old faces and old attributes, participants struggled to correctly pair these two items. One possible explanation is that the items are not bound together at encoding at higher set sizes, thus optimal recognition opportunities do not benefit associative memory (like only seeing old faces and old attributes). A second, possible explanation concerns recognition: The memory traces for the associations between faces and attributes are not strong enough to help the observer clearly remember which pair is correct. It may be that when the observer is presented with the old items (faces, and attributes), the memory traces for all these items activate, but no activation is higher than another. Since all the memory traces activate, the observer is unable to determine which pairing is correct.

Limitations

Despite analysing Hit Rate and False Alarm Rate for the three of types of test, signal detection measures could not be calculated. As mentioned previously, it was possible to calculate *d'* and *c* from the HR and FAR for faces and attributes, because these tests used a traditional one-interval design where one stimulus was presented at a time (Macmillan & Creelman, 2005), and the responses, 'Hits' and 'False Alarms' fell on a binomial distribution with equality probability for both outcomes (Lee & Wagenmakers, 2014). The third test for pairs was different: Participants were shown only old faces, and were given a list of all the old attributes from that trial. Their task was to decide which attribute from that list matched the face on the screen. The characteristics of this task are different from the tests used for faces and attributes. First, the binomial distribution is not equal. The probability of making a correct or incorrect decision is not 0.5, because the participant must choose which attribute from the list matches the face. Therefore, chance performance is set to 1/set size, and the probability of a Hit and False Alarm is 1/Set Size and (Set Size-1)/Set Size respectively. Furthermore, these probabilities changed per trials since the participant could use sampling with replacement. In my opinion, the most important characteristic that further complicated this task was that there was no way to measure decision threshold with this task. Participants were forced to make a decision, and were not allowed to reject the face. For this reason, it was not possible to determine the sensitivity of the decision threshold from this data. Stanislaw and Todorov express a similar view point about alternative forced choice tasks [mAFC], and argue that "because this comparison does not involve a criterion, mAFC[62] tasks are only suitable for measuring sensitivity" (Stanislaw & Todorov, 1999, pp 141). For the current experiment, sensitivity ranges between 1/M (where M indicates the number of options available) and 1 (perfect performance) and is calculated in the way that HR is calculated.

There are other limitations that arise from the paired associate test used in the current experiment. First, the paired associate test suffers from a degree-of-freedom problem, because decision-making is easier in the lower set sizes, since (a) there are fewer options to choose from, and (b) participants can make decisions by excluding the previously used attributes. For example, if a participant in Set Size 2 felt confident in their first decision, then they could easily make a decision for the second face presented. To combat this, the researcher can either (a) always provide the participants with more than one option to choose between (like a 2AFC test, but then criterion cannot be calculated) or (b) use a different type of task.

The second limitation is that larger set sizes have longer lists of old attributes, and these increase the task difficulty at recognition as argued by Podd (1990), Metzger (2002), and Lamont et al. (2005). Thus, for the current study, task difficulty at recognition and set size are confounded in higher set size groups.

Finally, the reason to use only old attributes and old faces for the cued matching task was to provide participants with optimal conditions for matching. A consequence of this was

[62] mAFC refers to Alternative Forced Choice (AFC). The letter 'm' refers to the number of choices available.

that the errors reported for the pairs test were not 'true' false alarms (although I have termed them as such for the analyses), since no *new* items were presented - participants were not differentiating between signal and noise (Macmillan & Creelman, 2005). Instead, these false alarms are better classified as pairing-errors, and participants could make as many pairing-errors as there were faces and items. In fact, there were 1830 possible combinations of faces and attributes for the Set Size 30 group.

Thus, in Experiment Two, which is reported next, the cued-matching task was replaced with an Old-New task. Participants saw congruent pairs that consisted of correctly paired old faces and old attributes, or incongruent pairs that consisted of an old face and old attribute that were both seen at encoding, but not together.

Face Recognition Experiment 2:
Testing Memory for Pairs Using an Old-New Task

Method

Design

This experiment used a repeated measures design with one between-subjects factor (set size) with five levels (4, 6, 8, 12, or 24 face-attribute pairs). Participants were randomly assigned to one of the five experimental set size groups that differed according to many faces were shown at encoding. These set sizes were chosen so that (a) they were divisors of the same number so that an equal number of data points could be collected from each participant (in this experiment, of 24), and (b) were an even number so that an equal number of congruent and incongruent pairs could be constructed. The repeated measures factor, test type, had three levels: faces, attributes, and pairs.

Sample

Sixty-seven participants were recruited from the undergraduate caucus of psychology students at the University of Cape Town; these participants had not participated in Experiment 1. Their average age was 21.3 years (*SD* = 4.47 years), and 59 participants (88.06%) were female.[63]

Due to missing data, the design was slightly unequal. In total, 13 participants were assigned to Set Size 4, Set Size 12, and Set Size 24 respectively. The remaining two groups – Set Size 6 and Set Size 8 - contained 14 participants each.

Materials

Faces at encoding. Twenty-Four old faces were randomly selected from the 30 0ld faces used in Face Recognition Experiment 1. Henceforth these faces are referred to as 'old', because they comprised the material shown at encoding.

Attributes at encoding. Twenty-Four attributes were randomly selected from the attribute list used in Face Recognition Experiment 1. These attributes are listed in Appendix I. These attributes are 'old' as they were shown at encoding.

Attributes at recognition. For Face Recognition Experiment 1, new attributes were generated to match the old attributes. The new attributes used in the current experiment were the same attributes generated to match the randomly-chosen old attributes used in Face Recognition Experiment 1, and are listed in Appendix I.

Order of old face-attributes pairs. Two random orders of face-attribute pairs were created to counterbalance any effects that could arise from the pairing and facilitate recognition, for example, a stimulus image that was perceived as 'innocent' or 'baby faced'

[63] Most participants (*n* = 37; 55.22%) identified as white South Africans, followed by equal numbers who identified as Black, or Coloured, (*n* = 12; 17.91% for each group), five participants (7.46%) identified as Indian, and 1 participant (1.49%) identified as Other.

would be considered a good match with the attribute 'He sings in a church choir'. The two pairs are listed in Appendix J. The first two digits of the names of both stimuli indicate the pairing, therefore 01 Face and 01 Fact were paired together.

Equal numbers of participants in Set Size 6 and Set Size 8 received stimuli paired using Order A and Order B. For Set Sizes 4 and 12, six and seven participants viewed stimuli pairs paired using Order A and B respectively; for Set Size 24, seven and six participants viewed stimuli pairs paired using Order A and B respectively.

Procedure

Participants came to the laboratory, where they signed a consent form before being seated at a computer. Participants were randomly assigned to one of five experimental set size groups. The experiment was split into three stages: encoding, filler task, and recognition.

Encoding

Participants were warned that they would view a series of faces and attributes that they had to study and recognise later. Each face appeared on the screen for three seconds, followed by a attribute that appeared beneath the face for three seconds, after which the face disappeared and the attribute remained on the screen for another three seconds. An interstimulus interval of 500 ms was used. The pairing of the faces and the attributes was predetermined, but the order in which the face-attribute pairs appeared within each trial was randomised.

Filler task

After encoding, participants completed the filler task where they were given a random puzzle. The allotted time to compete the puzzle varied between the set size groups. Like Face Recognition Experiment 1, the duration of the filler tasks differed to account for the delay between the onset of the median face at encoding and the onset of the recognition stage across set sizes. Thus, the encoding duration was longer for the larger set sizes, and subsequently the delay between the onset of the median face and the start of the recognition trial was controlled for across set sizes by adjusting the filter task duration. The filler task time ranged from 183 seconds to 88,5 seconds for Set Size 4 and Set Size 24 respectively.

Recognition

During the recognition stage, participants completed three Old-New tests for faces, attributes, and pairs. Before the recognition stage began, participants received a revised version of the Know and Remember instructions (Gardiner & Java, 1990; Yonelinas, 2001; see Appendix K). These instructions were revised to suit the stimuli used in the experiment.

For the face recognition test, participants were shown all the old faces from that trial and an equal number of new faces, one at a time in a randomised order, and had to indicate whether they had studied the face during that trial. The old and new faces were unique to each trial, and did not repeat across trials. Participants made their response via keypress by pressing '1' for Old, and '0' for New. After making their decision, participants were asked to rate their confidence in their decision on a scale that ranged from 1 - 2 - 3, with confidence rank indicating 'sure', 'less sure', and 'very unsure' respectively (Gardiner & Java, 1990; Yonelinas, 2001; Yonelinas & Parks, 2007). If participants had stated that they did not recognise the face (i.e., it was New), then they were instructed to respond 'N'; however, if they had indicated that they recognised the face they were encouraged to further indicate whether they 'Remember' or 'Know' the face (please see Appendix K for these instructions and definitions). After answering Know/Remember, participants moved to the next face within that trial, and repeated this procedure until they had responded to all the faces, before beginning the attribute recognition test.

The attribute recognition task followed the same procedure as the face recognition task: Participants were shown an equal number of old and new attributes, in a randomised order, one at a time, and had to indicate whether the attribute was old or new. Participants were tested on all the attributes that they had studied in that trial. After making their decision, they were prompted for a confidence rating using the same 1 - 2 - 3 response, and indicated if they had a Know-response or a Remember-response to that attribute. Participants were presented with the same definitions of ‘Know’ and ‘Remember’ before viewing the attributes. The attribute test continued until participants had responded to all the old and new attributes in that trial.

The final test was a face-attribute pairs recognition test using an Old-New task. Participants were shown either congruent (i.e., old), or incongruent (i.e., new) face-attribute pairs. Congruent pairs comprised the correctly paired old face and old attribute, whereas incongruent pairs comprised an incorrectly paired old face and old attribute. No new faces or attributes were shown during this stage, and only old faces and old attributes from that trial comprised the pairs. Half of the trials in this test were congruent pairs, and the other half were incongruent pairs. Each face and fact were used in only one pair combination. Before beginning this stage of the test, participants received the same reminder about the ‘Know’ and ‘Remember’ definitions. After this, participants were shown the congruent and incongruent pairs in randomized order, one pair at a time, and had to indicate whether the pair was old (i.e., congruent) or new (i.e., incongruent). Following each decision, participants rated their confidence on the 1-2-3 rating scale, and indicated if they experienced a Know-response or a Remember-response to the pairs. Participants made these responses to all the stimuli pairs for that trial. After this, participants started the next trial of the experiment until they had encoded and been tested on all 24 faces-attribute pairs. Participants in the Set Size 4 condition completed six trials, whereas participants in the Set Size 24 condition completed only one trial. Participants in the Set Size 6, 8, and 12 conditions completed four, three, and two trials respectively. The procedure is shown in Figure 4.11.

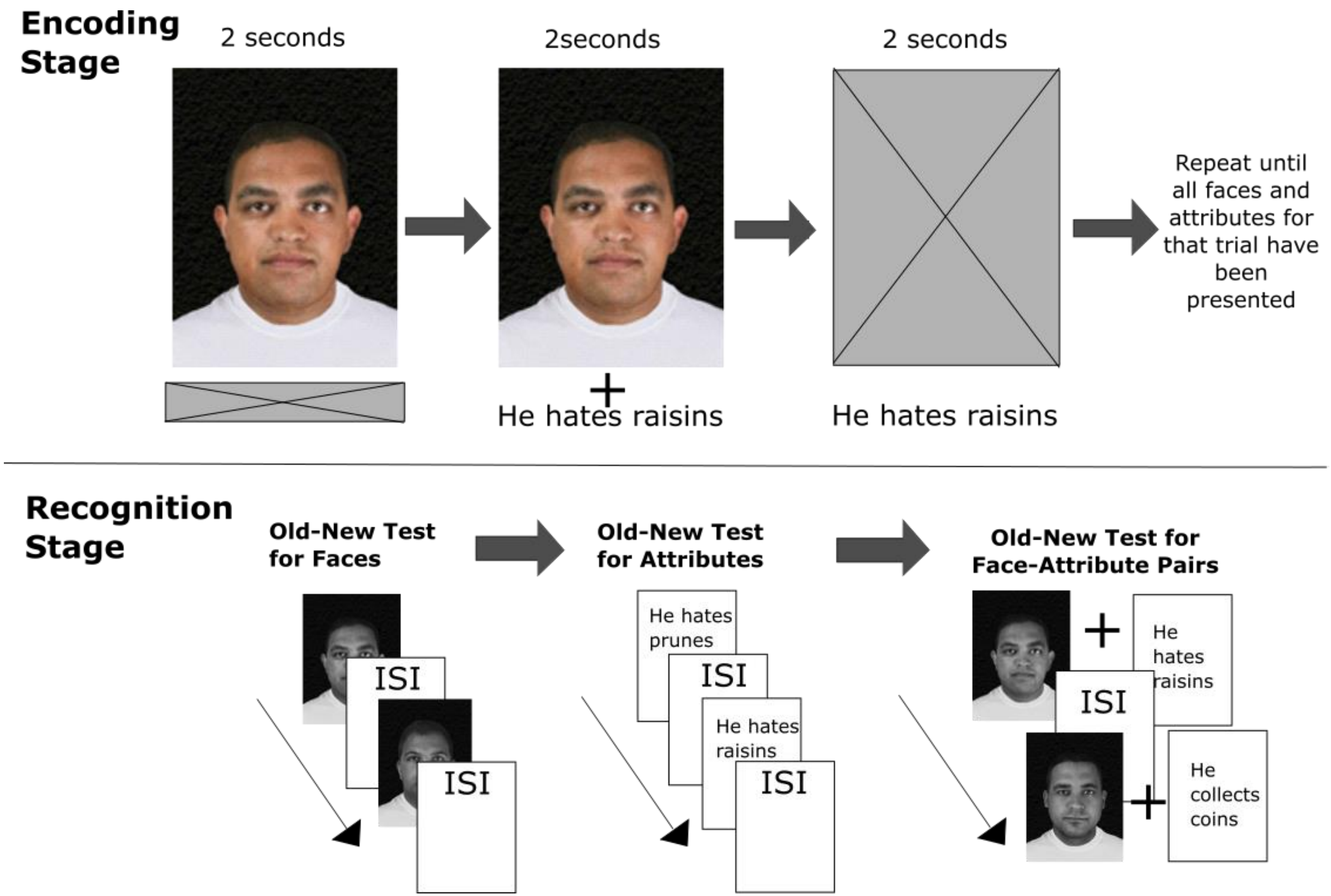

Figure 4.11. Illustration of the procedure used in Face Recognition Experiment 2. During the encoding stage, participants were presented with a face-attribute pair for a total of nine seconds. After a short filler task, participants completed three tests during the recognition stage: an Old-New test for Faces, an Old-New test for Attributes, and an Old-New test for Pairs.

Results

In this experiment, responses to all three tests (faces, attributes, and pairs) were coded as Hits or False Alarms, and from this an averaged Hit Rate (HR) and False Alarm Rate (FAR) were computed for each participant for each test. There were 24 old and 24 new trials for faces, and attributes, respectively, but only 12 old and new trials for pairs. The descriptive statistics for HR and FAR are reported in Table 4.13 and in Figure 4.12. Overall, the results suggest that recognition performance was highest for attributes. Recognition performance appears to be similar across faces and pairs.

Table 4.13

Descriptive Statistics for Hit Rate, False Alarm Rate Across Test Type and Set Size

	Faces		Attributes		Pairs	
Set Size	HR	FAR	HR	FAR	HR	FAR
4	0.72 (0.02)	0.21 (0.04)	0.93 (0.02)	0.03 (0.01)	0.70 (0.03)	0.12 (0.04)
6	0.67 (0.03)	0.15 (0.03)	0.94 (0.02)	0.02 (0.01)	0.65 (0.04)	0.17 (0.04)
8	0.60 (0.02)	0.14 (0.02)	0.95 (0.01)	0.05 (0.02)	0.52 (0.04)	0.20 (0.05)
12	0.80 (0.03)	0.25 (0.04)	0.94 (0.02)	0.06 (0.02)	0.81 (0.04)	0.29 (0.06)
24	0.69 (0.04)	0.26 (0.04)	0.89 (0.03)	0.12 (0.04)	0.67 (0.03)	0.37 (0.06)
Average	0.69 (0.01)	0.20 (0.02)	0.93 (0.01)	0.06 (0.01)	0.67 (0.02)	0.23 (0.02)

Note. HR and FAR denote Hit Rate and False Alarm Rate respectively. The number of old, and new trials for faces and attributes was 24 (so 48 in total), whereas the number of old, and new trials for pairs was 12 (so 24 in total). The groups are slightly uneven with one fewer participant in Set Size 4,6, and 24 (n =13 in each group) than in Set Size Group 8, and 12 (n = 14 in each group). Values in parentheses denote standard errors. Chance is 0.5.

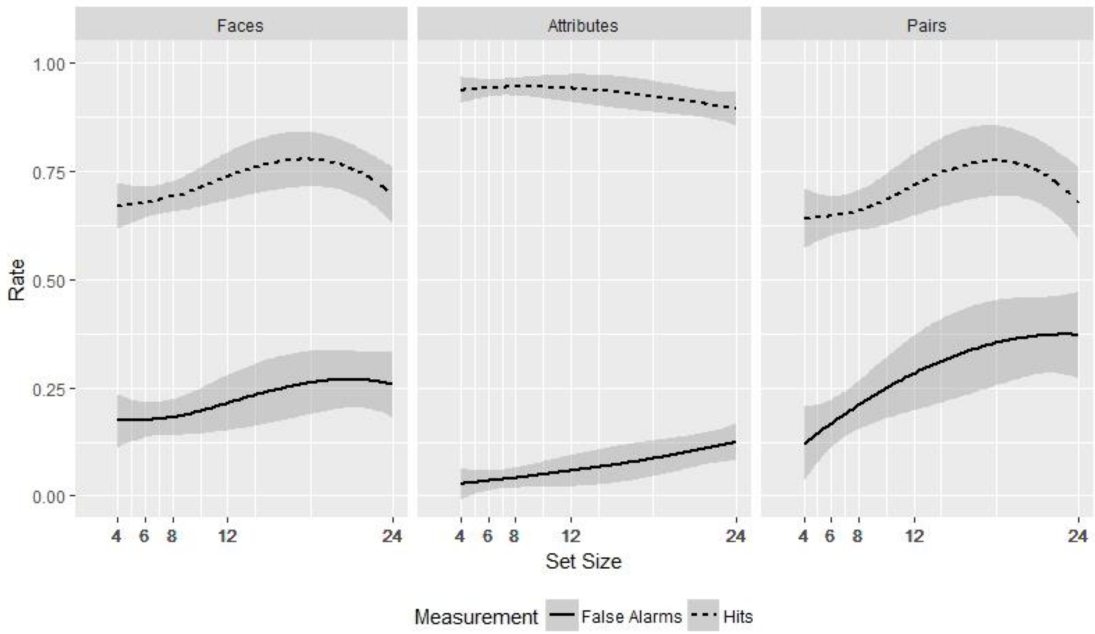

Figure 4.12. Hits and false alarms as a function of set size, for three types of tests. Chance is .50. The shaded areas are 95% confidence intervals. Curves are LOESS non-parametric, with span = 1.3.

Two signal detection measures, *d'* and *c*, were calculated from the HR and FAR. The descriptive data for the signal detection measures are reported in Table 4.14.

Table 4.14
Descriptive Statistics for d' and c Across Test Type and Set Size

	Faces		Attributes		Pairs	
Set Size	*d'*	*c*	*d'*	*c*	*d'*	*c*
4	1.55 (0.18)	0.18 (0.09)	3.36 (0.14)	0.11 (0.07)	1.80 (0.17)	0.35 (0.09)
6	1.57 (0.14)	0.34 (0.09)	3.51 (0.15)	0.10 (0.06)	1.44 (0.20)	0.31 (0.08)
8	1.44 (0.15)	0.47 (0.06)	3.38 (0.17)	0.02 (0.08)	0.99 (0.18)	0.45 (0.11)
12	1.69 (0.17)	-0.05 (0.11)	3.24 (0.18)	0.04 (0.06)	1.61 (0.21)	-0.14 (0.12)
24	1.31 (0.15)	0.09 (0.13)	2.63 (0.22)	-0.03 (0.08)	0.88 (0.23)	-0.05 (0.10)
Average	1.51 (0.07)	0.21 (0.05)	3.23 (0.08)	0.05 (0.03)	1.34 (0.1)	0.19 (0.05)

Note. The number of old and new trials for faces and attributes was 24 (so 48 in total), whereas the number of old and new trials for pairs was 12 (so 24 in total). The groups are slightly uneven with one fewer participant in Set Size 4, 6, and 24 (*n* =13 in each group) than in Set Size 8 and 12 (*n* = 14 in each group). Values in parentheses denote standard errors. Both *d'* and *c* were calculated using the correction to extreme values as outlined in Stanislaw & Todorov (1999).

Mixed Linear Model for D-Prime

Hit Rate and False Alarm Rate were not analysed separately, because their relationship with each other was incorporated into the signal detection measure, *d'*. Two separate mixed linear models were used to fit the remaining two dependent variables, *d'* and *c*. Each model will be discussed in turn. These models were fit and analysed using the `lme4` package (Bates et al., 2015), because random effects were included in the models. In each model, *d'* and *c* were nested within subject.

First, a null model was constructed where *d'* was predicted by the intercept and random effects for each subject. A second model was constructed where *d'* was predicted by fixed effects, set size and test type, with random effects for subject. Finally, a third, full model was constructed where *d'* was predicted by fixed effects, set size and test type, and an interaction term between set size and test type, with random effects for subject. The output, model comparisons, and model syntax are listed in Table 4.15. The second model was a better fit than the null model, χ^2 (6) = 237.94, $p < .001$, and the third model was a better fit than the second model, χ^2 (8) = 20.066, $p = .010$. The third model accounted for between 68% and 81 % of the variance (R_m^2 = 0.68 and R_c^2 = 0.81). All subsequent analyses were performed on Model 3.

Table 4.15

Characteristics and Chi-Square Tests of the Three Models Predicting d'.

Model Name	Model details	df	AIC	BIC	Log likelihood	Deviance	$\Delta\chi^2$	R_m^2	R_c^2
Null Model	d' ~ 1 + (1 | Subject)	3	613.10	623.01	-303.55	607.10		0	0
Model 2	d'~ Set size + Test + (1 | Subject)	9	387.16	416.89	-184.58	369.16	**237.94**	0.66	0.73
Model 3	d'~ Set size *Test + (1 | Subject)	17	383.10	439.25	-174.55	349.10	**20.066**	0.68	0.81

Note. This table shows the characteristics of the three different models constructed in R. The chi-square values, and their respective p-values, are based on a comparison between that model and the previous model. There are only two comparisons: the null model versus model 2, and model 2 versus model 3. The asterisk in the model details for Model 3 indicates that Hits is predicted by both variables, and their interaction term.

Embolden text denotes significance ($p < .05$). The df for the chi-square statistic is the difference in the df of the two models.

The column 'Model details' lists the syntax for each model. The tilde symbol (~) denotes prediction, the number one '1' denotes the intercept and '1|' indicates random effects (i.e., intercepts) grouped by the subsequent term. Therefore, for the null model, the outcome variable *d'* is predicted by the intercept, and by random effects for each subject.

A closer examination of the Model 3 using the `car` package (Fox & Weisberg, 2011) showed that there was a significant main effect for set size, $\chi^2(4) = 14.907$, $p = .005$, a significant main effect for test type, $\chi^2(2) = 651.39$, $p < .001$, and a significant interaction between set size and test type, $\chi^2(8) = 21.646$, $p = .006$. The coefficients of the fixed effects are in Table 4.16.

Table 4.16.

Model Coefficients for Model 3 Predicting d'

Fixed Effects	B	CI lower	CI upper	β	CI lower	CI upper	p
Intercept (Set Size 4, Attributes Test)	3.36	3.02	3.70	-	-	-	<.001
SetSize 6	0.15	-0.32	0.63	0.06	-0.12	0.23	.523
SetSize 8	0.02	-0.45	0.49	0.01	-0.17	0.18	.934
SetSize 12	-0.12	-0.60	0.37	-0.04	-0.22	0.13	.639
SetSize 24	-0.72	-1.21	-0.24	-0.26	-0.44	-0.09	**.004**
Faces Test	-1.81	-2.18	-1.45	-0.78	-0.94	-0.62	**<.001**
Pairs Test	-1.56	-1.93	-1.20	-0.67	-0.83	-0.52	**<.001**
Set Size6: Faces Test	-0.13	-0.63	0.38	-0.03	-0.15	0.09	.626
SetSize8: Faces Test	-0.12	-0.63	0.38	-0.03	-0.15	0.09	.630
SetSize12: Faces Test	0.26	-0.25	0.78	0.06	-0.06	0.17	.322
SetSize24: Faces Test	0.49	-0.03	1.00	0.11	-0.01	0.23	.670
SetSize6: Pairs Test	-0.52	-1.02	-0.01	-0.12	-0.24	0.00	.480
SetSize8: Pairs Test	-0.82	-1.33	-0.32	-0.19	-0.31	-0.07	**.002**
SetSize12: Pairs Test	-0.07	-0.59	0.44	-0.02	-0.13	0.10	.783
SetSize24: Pairs Test	-0.20	-0.71	0.32	-0.04	-0.16	0.07	.456
Random Effects							
Participants							
Number of observations	201						
Number of participants	67						
ICC	0.423						
Variance	0.226						

Note. Embolden text indicates significance at alpha = .05.

The main effect of set size was best predicted by a linear trend, $t(67) = -2.912$, $p = .005$. Consecutive planned contrasts showed no significant difference between Set Sizes 4 and 6, Set Sizes 6 and 8, and Set Sizes 8 and 12, all *ps* > .05. There was, however, a significant difference between Set Sizes 12 and 24, $t(67) = -2.992$, $p = .015$ (see Figure 4.13). Participants in Set Size

24 obtained a significantly lower *d'* than those in Set Size 12, $d'_{Set\ size\ 24}$= 1.60, 95% CI [1.44, 1.77], SE = 0.17 versus $d'_{Set\ size\ 12}$= 2.18, 95% CI [2.02, 2.34], SE = 0.16.

For the significant main effect of test type, pairwise comparisons showed that *d'* was significantly higher for attributes than faces, $d'_{Attributes}$ = 3.23, 95% CI [3.14,3.32], SE = 0.08, versus d'_{Faces}= 1.51, 95% CI [1.44,1.58], SE = 0.07, $t(134) = 20.873$, $p < .001$, and significantly higher for attributes than pairs, $d'_{Attributes}$= 3.23, 95% CI [3.14,3.32], SE = 0.08, versus d'_{Pairs}= 1.34, 95% CI [1.24,1.44], SE = 0.09, $t(134) = 22.943$, $p < .001$. There was no significant difference between *d'* for faces versus pairs, $t(134) = 2.070$, $p = .10$. These results are shown in Figure 4.13.

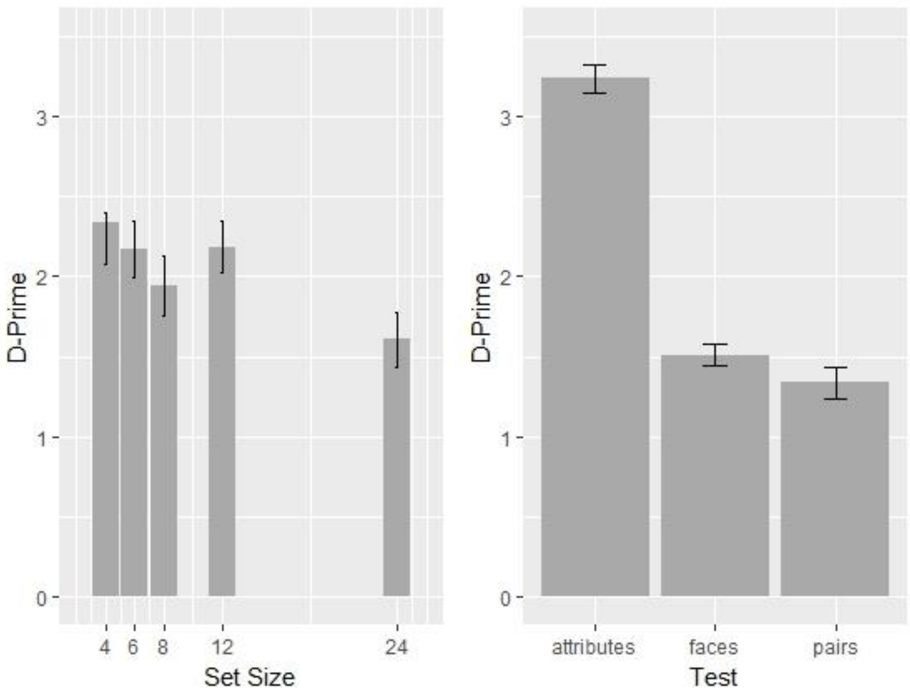

Figure 4.13. Marginal means for *d'* across the set size and test type. The error bars represent 95% confidence intervals. P-values were adjusted for multiple comparisons.

There was a significant interaction between set size and test type. The means for each test, within each Set Size are shown in Figure 4.14, and pairwise comparisons are listed in

Table 4.17. Inspection of Figure 4.14 and Table 4.17 show that *d'* was consistently significantly higher for attributes than pairs and faces, which did not differ from each other.

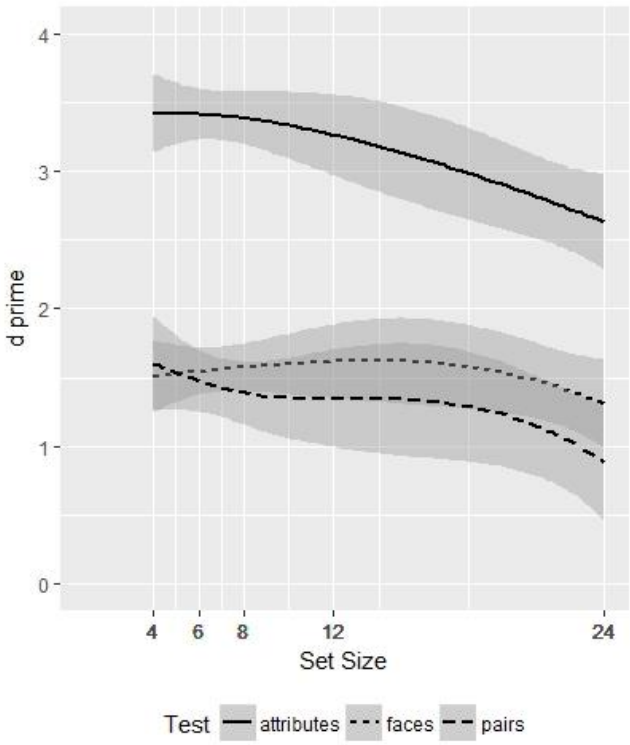

Figure 4.14. Mean *d'* across Set Sizes (4, 6, 8, 12, 24) for three tests, faces, attributes, and pairs. The shaded areas are 95% confidence intervals. Curves are LOESS non-parametric, *span = 1.3*. The solid black line represents the results for the attributes test, the dashed line shows the results for the faces test, and the long-dash line represents the results for the pairs test.

Table 4.17

Pairwise Comparisons Between Test Type Within Set Size

Set Size	Test	*t*	df	p
4	Attributes vs Faces	9.735	134.00	**<.0001**
	Attributes vs Pairs	8.385	134.00	**<.0001**
	Pairs vs Faces	-1.350	134.00	.990
6	Attributes vs Faces	10.807	134.00	**<.0001**
	Attributes vs Pairs	11.572	134.00	**<.0001**
	Pairs vs Faces	0.765	134.00	1.000
8	Attributes vs Faces	10.798	134.00	**<.0001**
	Attributes vs Pairs	13.293	134.00	**<.0001**
	Pairs vs Faces	2.496	134.00	.451
12	Attributes vs Faces	8.329	134.00	**<.0001**
	Attributes vs Pairs	8.776	134.00	**<.0001**
	Pairs vs Faces	0.447	134.00	1.000
24	Attributes vs Faces	7.120	134.00	**<.0001**
	Attributes vs Pairs	9.443	134.00	**<.0001**
	Pairs vs Faces	2.323	134.00	.576

Notes. Embolden text indicates p-values significant at .05. P-values were adjusted for multiple pairwise comparisons.

Mixed Linear Model for Criterion

A second mixed linear model was used to predict criterion. Three mixed linear models were constructed. In the first model – i.e., the Null model – *c* was predicted by the intercept, and random intercepts for each subject. A second model was built where *c* was predicted from fixed effects, set size and test type, and the random effects for subject. In the third model, a third fixed effect, the interaction term between set size and test type, was added as a predictor to the two predictors and the random effects for subject used in Model 2. Each model was compared to the subsequent model. The second model was a significant improvement over the null model, $\chi^2(6) = 33.608$, $p < .001$. Model 3 was a significant improvement over Model 2, χ^2 (8) = 24.700, $p = .002$. Model 3 accounted for between 26% and 47% of the variance ($R^2_{marginal}$ = 0.26 and $R^2_{conditional}$ = 0.47). These results are reported in Table 4.18. All subsequent analyses were performed on Model 3.

Table 4.18

Characteristics, and Chi-Square Tests of the Three Models Considered for c

Model Name	Model details	df	AIC	BIC	Log likelihood	Deviance	$\Delta\chi^2$	R_m^2	R_c^2
Null Model	c ~ 1 +(1 | Subject)	3	159.06	168.96	-76.527	153.053		0	0
Model 2	c~ Set size + Test + (1 | Subject)	9	137.44	167.18	-59.723	119.445	**33.608**	0.19	0.36
Model 3	c~ Set size *Test + (1 | Subject)	17	128.75	184.90	-47.372	94.745	**24.700**	0.26	0.47

Note. This table shows the characteristics of the three different models predicting *c*. The chi-square values, and their respective p-values, are based on a comparison between that model and the one before it. There are only two comparisons: the null model versus model 2, and model 2 versus model 3. The asterisk in the model details for Model 3 indicates that Hits is predicted by both variables, and their interaction term.
Embolden text denotes significance (p < .05). The df for the chi-square statistic is the difference in the df of the two models.
The column ‘Model details’ lists the syntax for each model. The tilde symbol (~) denotes prediction, the number one ‘1’ denotes the intercept and ‘1|’ indicates random effects (i.e., intercepts) grouped by the subsequent term. Therefore, for the null model, the outcome variable *c* is predicted by the intercept, and by random effects for each subject.

There was a significant main effect of set size, $\chi^2(4) = 25.944$, p <.001, and of test type, $\chi^2(2) = 14.673$, $p < .001$. There was also a significant interaction between set size and test type on criterion, $\chi^2(8) = 27.123$, $p < .001$. The coefficients for the fixed effects are presented in Table 4.19

Table 4.19

Model Coefficients for Model 3 Predicting c

Fixed Effects	B	CI lower	CI upper	β	CI lower	CI upper	p
Intercept (Set Size 4, Attributes Test)	0.11	-0.06	0.28	-	-	-	.206
SetSize 6	-0.01	-0.25	0.23	-0.01	-0.28	0.25	.931
SetSize 8	-0.09	-0.33	0.15	-0.10	-0.37	0.16	.443
SetSize 12	-0.07	-0.32	0.17	-0.08	-0.34	0.19	.248
SetSize 24	-0.14	-0.39	0.10	-0.15	-0.42	0.11	.248
Faces Test	0.07	-0.14	0.28	0.09	-0.17	0.35	.503
Pairs Test	0.24	0.03	0.45	0.31	0.04	0.57	**.025**
Set Size6: Faces Test	0.17	-0.12	0.46	0.12	-0.08	0.32	.250
SetSize8: Faces Test	0.38	0.09	0.66	0.26	0.06	0.46	**.011**
SetSize12: Faces Test	-0.16	-0.46	0.13	-0.11	-0.30	0.09	.277
SetSize24: Faces Test	0.05	-0.25	0.34	0.03	-0.16	0.23	.751
SetSize6: Pairs Test	-0.03	-0.32	0.26	-0.02	-0.22	0.18	.841
SetSize8: Pairs Test	0.19	-0.09	0.48	0.13	-0.06	0.33	.188
SetSize12: Pairs Test	-0.42	-0.71	-0.12	-0.28	-0.47	-0.08	**.006**
SetSize24: Pairs Test	-0.25	-0.55	0.04	-0.17	-0.36	0.03	.093
Random Effects							
Participants							
Number of observations	201						
Number of participants	67						
ICC	0.281						
Variance	0.072						

Note. ICC denotes intraclass correlation. Embolden text indicates significance at alpha = .05.

The main effect of test was further examined. Pairwise comparisons showed a significant difference in criterion between attributes and faces, $t(134) = -3.375$, $p = .003$. There was also a significant difference in criterion between attributes and pairs, $t(134) = -2.983$, $p < .001$. There was no significant difference between faces and pairs, $t(134) = 0.392$, $p = .919$. Participants adopted a more conservative criterion for faces compared to pairs, c_{faces}= 0.211, 95% CI [0.163, 0.259], SE =0.048, versus c_{pairs}= 0.192, 95% CI [0.140, 0.244], SE =0.052, and attributes, $c_{attributes}$= 0.048, 95% CI [0.017,0.079], SE = 0.031.

The significant main effect of set size was best fit with a linear trend, $t(67) = -3.625$, $p < .001$, cubic trend, $t(67) = 1.963$, $p = .054$, or quartic trend, $t(67) = 2.469$, $p = .016$. Consecutive comparisons showed that there were no significant differences between criterion at Set Size 4 and 6, 6 and 8, or 12 and 24. The only significant difference in criterion was between Set Sizes 8 and 12, $t(67) = -4.086$, $p <.001$. These results are shown in Figure 4.15.

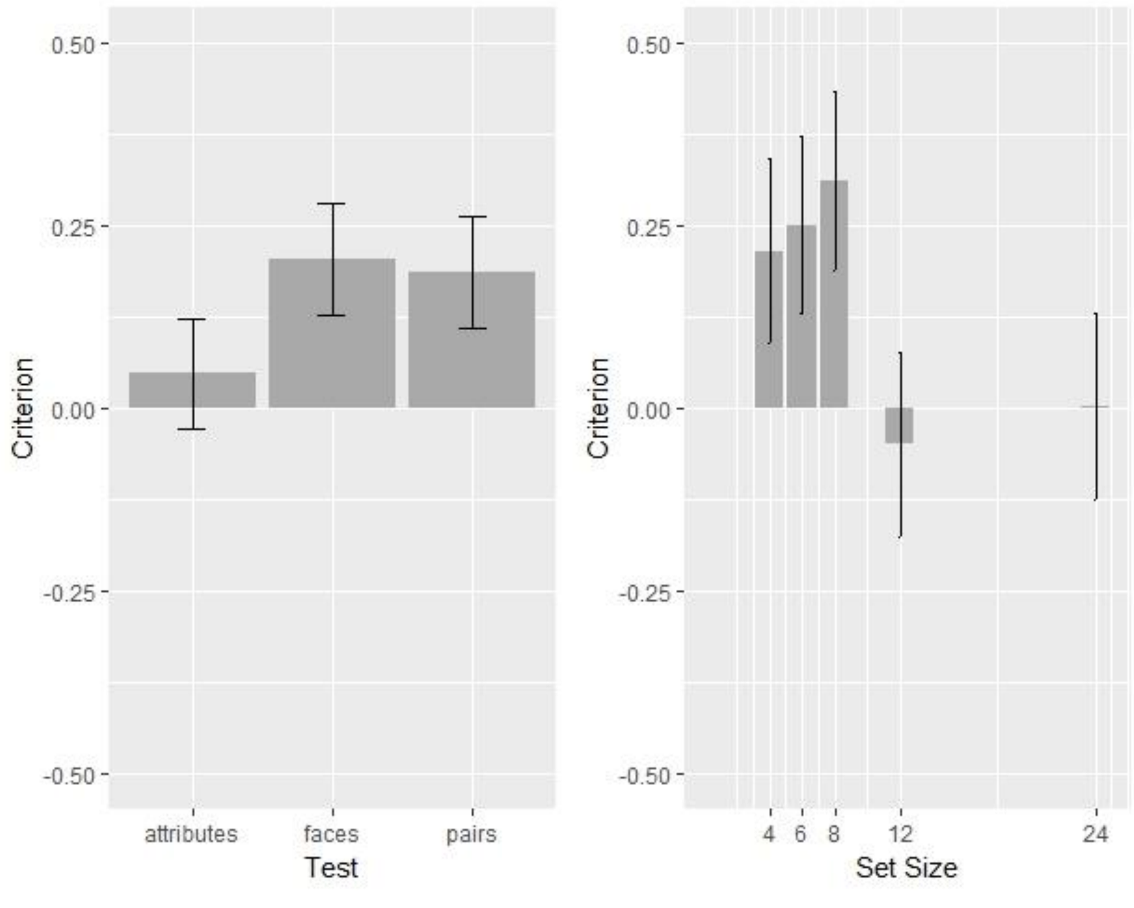

Figure 4.15. Marginal means for *c* across test type and set size. The image on the left shows the marginal means for *c* across the three types of test. The image on the right shows the marginal means for *c* across set size. The error bars represent 95% confidence intervals.

Pairwise comparisons to investigate the significant interaction between set size and test type showed that at Set Size 8, criterion was significantly different between attributes and faces, $t(134.00) = -4.399$, $p = .002$, and between attributes and pairs, $t(134.00) = -4.263$, $p = .003$. At

Set Size 8, participants adopted a more conservative criterion for faces, and pairs, compared to attributes. There were no other significant pairwise comparisons. These results are shown in Figure 4.16, and the pairwise comparisons are listed in Table 4.20.

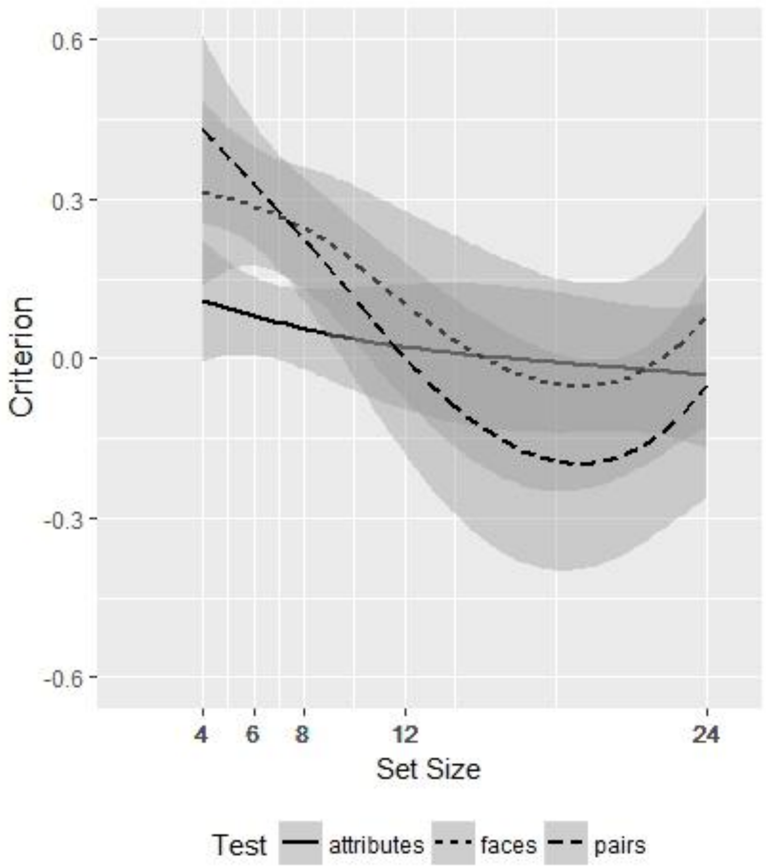

Figure 4.16. Mean *c* across Set Sizes (4, 6, 8, 12, 24) for three tests, faces, attributes, and pairs. The shaded areas are 95% confidence intervals. Curves are LOESS non-parametric, *span = 1.3*. The solid black line represents the results for the attributes test, the dashed line shows the results for the faces test, and the long-dash line represents the results for the pairs test.

Table 4.20

Pairwise Comparisons Between c of Levels of Test Type Within Set Size

Set Size	Test	t	df	p
4	Attributes vs Faces	-0.671	134.00	1.000
	Attributes vs Pairs	-2.271	134.00	.613
	Pairs vs Faces	-1.600	134.00	.956
6	Attributes vs Faces	-2.363	134.00	.546
	Attributes vs Pairs	-2.068	134.00	.753
	Pairs vs Faces	.295	134.00	1.000
8	Attributes vs Faces	-4.399	134.00	**.002**
	Attributes vs Pairs	-4.263	134.00	**.003**
	Pairs vs Faces	0.136	134.00	1.000
12	Attributes vs Faces	.871	134.00	.999
	Attributes vs Pairs	1.677	134.00	.936
	Pairs vs Faces	0.807	134.00	1.000
24	Attributes vs Faces	-1.121	134.00	.999
	Attributes vs Pairs	0.121	134.00	1.000
	Pairs vs Faces	1.242	134.00	.999

Note. Embolden text indicates p-values significant at .05. P-values have been adjusted to account for multiple pairwise comparisons.

ROC Curves

Receive operator characteristic (ROC) curves were constructed for the various set sizes across the three test types. To do this, confidence scores were transformed and ranked according to the following procedure described in Yonelinas and Parks (2007): Remember responses paired with a confidence score of ‘1’ (i.e., ‘Sure’), ‘2’, and ‘3’ were coded as ‘6’, ‘5’, and ‘4’ respectively, and Know responses, paired with a confidence score of ‘1’ (i.e., ‘Sure), ‘2’, and ‘3’ were coded as ‘3’,’2’, and ‘1’ respectively. When these transformed confidence scores were ranked, they were in order of highest to lowest confidence.

Once the data was transformed, an ROC curve was plotted for each test type following the tutorial created by Laura Mickes.[64] The cumulated Hits and False Alarms were calculated for each set size, and plotted where each point represents the HR and FAR at that confidence level. There were six points on each ROC curve. The diagonal reference line in each graph

[64] http://www.mickeslab.com/handy/roc-tutorial/

indicates equal HR and FAR, i.e., chance performance. An ROC curve that demonstrates better discriminability across various confidence levels moves away from the diagonal line towards the top left corner of the cartesian place: This indicates better diagnosticity since a high HR is coupled with low FAR.

I used the `pROC` package (Robin et al., 2011) to compare ROC curves. ROC curves are compared to each other by comparing their respective area under the curve (AUC). In situations where the FAR is truncated and does not extend to 1, the AUC is replaced by the partial area under the curve (pAUC). The pAUC can also be used to compare different ROC curves that extend to different points on the x-axis, thus covering different areas. A larger pAUC indicates better diagnosticity. In situations where two pAUC are compared, the researcher must decide whether (a) to only compare the overlapping areas (this is normally a more conservative area), or (b) to 'extend' the shorter curve to the same length of the second, longer curve, and then compare them. Following the recommendations by Gronlund, Wixted, and Mickes (2014), I chose the more conservative method.

When comparing two ROC curves to each other, a *Z* statistic is calculated from the difference between the curves using the formula $z = \frac{\theta_1 - \theta_2}{sd(\theta_1 - \theta_2)}$ where θ_1 and θ_2 denotes (p)AUC for ROC curve 1 and ROC curve 2 respectively. The `pROC` package uses bootstrapping (*n*=2000) when calculating the standard deviation of the differences between the two (p) AUC: For each iteration, the ROC curves, (p)AUC, and differences between the two (p)AUC are calculated, and from this, the standard deviation is calculated. Currently the `pROC` package only allows comparison of two ROC curves; multiple pairwise comparisons can be calculated, but the reported p-values will not be adjusted and would need to be adjusted for multiple comparisons separately. As mentioned, if comparing pAUC of two ROC curves with each other, the researcher must decide whether to use a conservative or lenient approach. When the researcher uses the conservative approach, the two pAUCs are only compared across the overlapping area, which is defined by the smaller specificity (1-FAR) of the two curves. For this reason, the difference is calculated not between the two pAUC of the two respective ROC curves, but, instead, is calculated between the truncated pAUC and the second pAUC, and based on the bootstrapped sample estimates. Furthermore, if the researcher conducts multiple pairwise comparisons, the smaller specificity region must be specified for each comparison.

ROCs were constructed for each set size for each test type, and these are shown in Figure 4.17. Each ROC, and its respective analyses will be discussed according to test type.

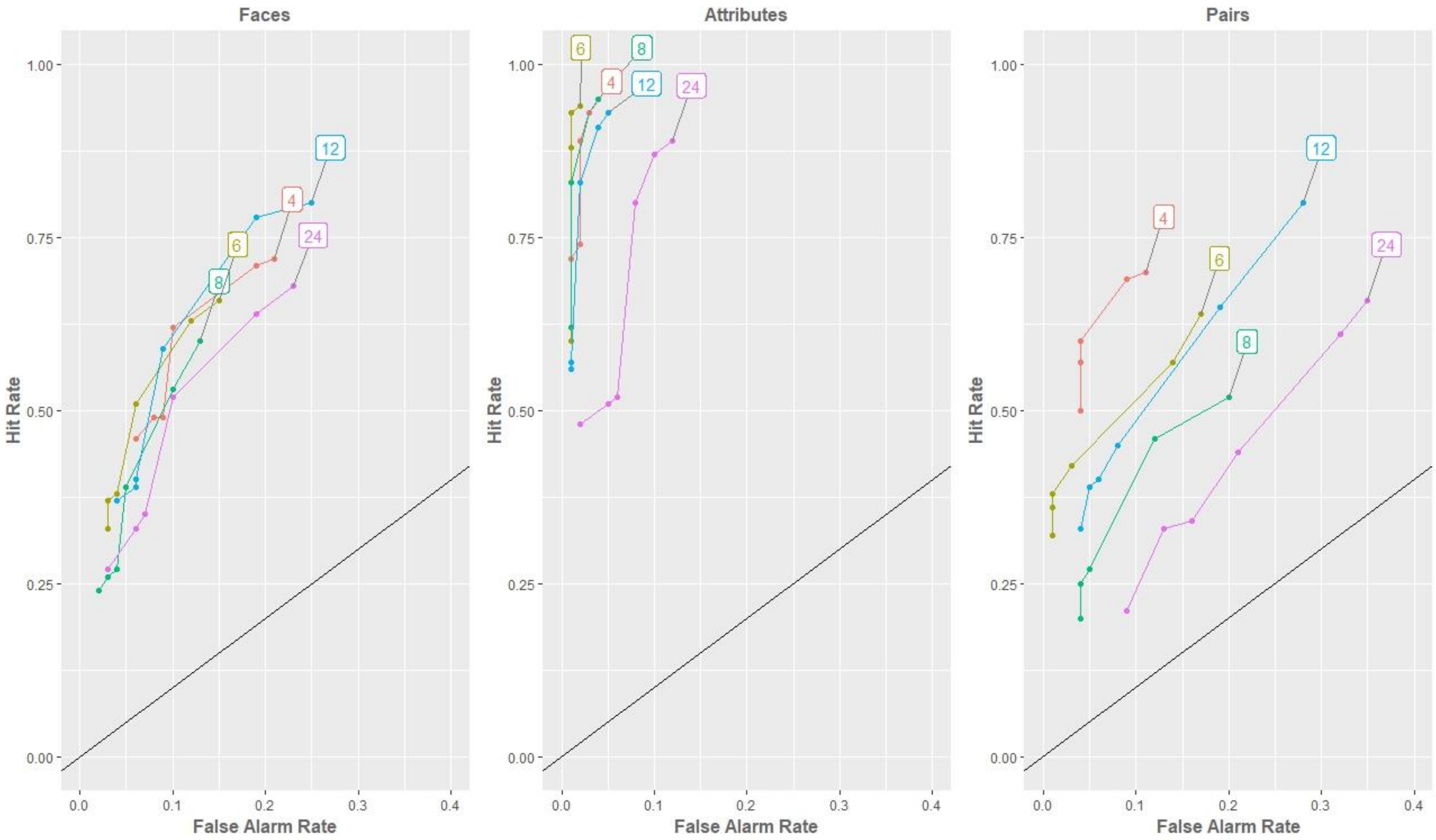

Figure 4.17. ROC curves for the three tests, faces, attributes, and pairs across Set Sizes 4, 6, 8, 12 and 24. Each ROC is labelled with its respective set size. The figure is scaled to the same FAR.

ROC curves for face recognition. ROC curves for face recognition were plotted across each set size and are shown in Figure 4.17. Since the results from the mixed linear model showed that the main effect of set size on *d'* was best fit with linear trend, and that the levels of set size had an ordinal nature, pAUC between set sizes were compared with adjacent contrasts, instead of pairwise comparisons. Therefore, the pAUC of each set size was compared to the pAUC of the subsequent set size. The results are listed in Table 4.21.

Table 4.21

Characteristics Comparisons of Each ROC Curve Within Set Size for Faces

Experimental Groups	pAUC	Lower CI	Upper CI	Specificity	Θ1 - Θ2	Θ1 Sample Estimate	Θ2 Sample Estimate	D	*p*
SS 4	.065	.056	.073	.792	SS 4 v. SS 6	.065	.067	-0.284	1
SS 6	.044	.038	.050	.845	SS 6 v. SS 8	.044	.041	0.835	1
SS 8	.034	.029	.039	.863	SS 8 v. SS 12	.079	.085	0.898	1
SS 12	.085	.075	.095	.750	SS 12 v. SS 24	.088	.087	1.538	.496
SS 24	.077	.067	.087	.744	-	-	-	-	-

Notes. D refers to the *Z* statistic calculated to compare pAUC areas, SS refers to Set Size, and Θ refers to the pAUC of the specific set size ROC curves. P-values are two-tailed, adjusted for four comparisons. Each set size ROC curve was compared to each subsequent ROC curve.

The results from Table 4.21 showed that there were no significant differences between the partial areas under the curve predicted by set size for the face recognition task. This result may appear surprising, because the results of the MLM show a significant main effect for set size, but the ROC curves and the MLM report different pairwise comparisons. For the ROC curve analyses, the effect of set size was compared within each test type (e.g., within faces, pAUC for various set sizes were compared to each other), whereas for the MLM, the three test types were compared within each set size (e.g., within Set Size 4, the three levels of test type were compared to each other). Pairwise comparisons between each Set Size within the three levels of test type were not calculated in the MLM. Thus, the ROC curves reported in this chapter provide further insight into the relationship between HR, FAR, and confidence that was not revealed using the MLM.

ROC curves for attribute recognition. ROC curves for the attributes were compared in the same way as for the faces: consecutive comparisons were conducted where the pAUC value for each set size ROC curve was compared with the subsequent set size ROC curve. The results are listed in Table 4.22.

The results from Table 4.22 show that the only significant difference was between the pAUC for Set Size 12 and Size 24 respectively. Specifically, the sample estimate for the pAUC that corresponded with the ROC curve for Set Size 12 was greater than the sample estimate for the pAUC that corresponded with the ROC curve for Set Size 24. The result suggests that participants displayed better diagnosticity in the Set Size 12 group than the Set Size 24 group.

Table 4.22

Characteristics Comparisons of Each ROC Curve Within Set Size for Attributes

Experiment Groups	pAUC	Lower CI	Upper CI	Specificity	Θ1 - Θ2	Θ1 Sample Estimate	Θ2 Sample Estimate	D	p
SS 4	0.013	0.011	0.015	0.968	SS 4 v. SS 6	0.013	0.014	-0.433	1
SS 6	0.009	0.008	0.011	0.976	SS 6 v. SS 8	0.024	0.022	0.971	1
SS 8	0.022	0.019	0.025	0.946	SS 8 v. SS 12	0.023	0.021	0.598	1
SS 12	0.021	0.018	0.024	0.946	SS 12 v. SS 24	0.056	0.046	2.656	**.032**
SS 24	0.046	0.040	0.051	0.875	-	-	-	-	-

Note. D refers to the Z statistic calculated to compare pAUC areas, SS refers to Set Size, and Θ refers to the pAUC of the specific Set Size ROC curves. P-values are two-tailed, adjusted for four comparisons. Each Set Size ROC curve was compared to each subsequent ROC curve. Embolden text indicates p-values significant at .05.

ROC curves for paired face-attributes recognition test. ROC curves were plotted for each set size within the pairs recognition test. These curves are presented in Figure 4.17. Unlike the ROC curves that were constructed for faces, and attributes, the ROC curves for face-attribute pairs showed greater separation between each Set Size. The pAUC for set sizes were compared with consecutive contrasts: The pAUC of each set size was compared with the pAUC of the subsequent set size. The results are listed in Table 4.23. Only one significant difference was found, and this was between the estimated pAUC for the ROC curve for Set Size 12 and the estimated pAUC for the ROC curve for Set Size 24. This result suggests that participants in Set Size 12 showed greater diagnosticity than participants in Set Size 24.

Table 4.23

Characteristics Comparisons of Each ROC Curve Within Set Size for Face-Attribute Pairs

Experiment groups	pAUC	Lower CI	Upper CI	Specificity	Θ1 - Θ2	Θ1 Sample Estimate	Θ2 Sample Estimate	D	p
SS 4	0.108	0.089	0.127	0.878	SS 4 v. SS 6	0.059	0.050	1.382	.667
SS 6	0.102	0.087	0.119	0.827	SS 6 v. SS 8	0.062	0.048	1.935	.194
SS 8	0.047	0.039	0.058	0.798	SS 8 v. SS 12	0.080	0.102	-2.053	.160
SS 12	0.050	0.041	0.059	0.711	SS 12 v. SS 24	0.147	0.108	2.715	**.027**
SS 24	0.039	0.031	0.046	0.628	-	-	-	-	-

Notes. D refers to the Z statistic calculated to compare pAUC areas, SS refers to Set Size, and Θ refers to the pAUC of the specific Set Size ROC curves. P-values are two-tailed, adjusted for four comparisons. Each Set Size ROC curve was compared to each subsequent ROC curve. Embolden text indicates p-values significant at .05.

Discussion

In the two experiments presented in this chapter, I tested the effect of set size on face recognition and associated attributes. In Experiment 1, participants were shown 1, 2, 3, 5, 10, 15 or 30 face-attribute pairs, and were subsequently tested using an Old-New task on faces and attributes (separately), and a cued-matching task on the association between faces and attributes. In Experiment 2, participants were shown 4, 6, 8, 12 or 24 face-attribute pairs, and like Experiment 1, were tested on their memory for the faces and attributes separately using an Old-New Task. Unlike Experiment 1 where associative memory was tested with a cued-matching task, in Experiment 2 associative memory was tested with an Old-New task comprising Old and New pairs.

Overall, the results showed that set size had a detrimental effect on recognition performance. In Experiment 1, both Hits and False Alarms decreased and increased, respectively, as set size increased. Recognition performance was worse for the cued-matching task for pairs than for Old-New task for faces and attributes respectively; in fact, the results clearly demonstrated that as set size increased, participants were unable to accurately pair faces and attributes despite demonstrating memory for these two items independently. Confidence was also negatively impacted by an increase in set size: Average confidence was lower for higher set sizes, and worst for the cued-matching task.

For Experiment 2, *d'* was negatively affected by set size, and was not uniform across test type. Significant differences in *d'* were found only between the higher Set Sizes, 12 and 24. As seen in Experiment 1, recognition performance was best for Attributes: *d'* was highest for attributes, but there were no significant differences in *d'* between faces and pairs. Furthermore, there was a significant interaction between test type and set size: *d'* was significantly highest for attributes across all set sizes, with no significant differences in *d'* between pairs and faces. Criterion was also affected by set size and test type. Participants adopted a more lenient criterion for attributes (with no differences in the criterion adopted for faces and pairs), and that the criterion adopted by participants changed from conservative to lenient as set size increased.

The ROC curves provided further insight into the relationship between HR and FAR, at various confidence levels, across the five set sizes within each test type. Comparisons of the pAUC for ROC showed no significant difference in diagnosticity among set size groups for faces; however, participants showed worse diagnosticity for Set Size 24 versus Set Size 12 for attributes and pairs, respectively. These results may appear contradictory to the findings from the MLM, since recognition performance was best for attributes compared to the faces and pairs. Why should there be a difference in diagnosticity between Set Sizes 24 and 12 for attributes, but no difference for faces? There are a few possible explanations: First, the MLM did not consider the differences between set sizes within each test group, that is, pairwise comparisons for *d'* between Set Size 4, 6, 8, 12 and 24 for faces were not examined. Instead, pairwise comparisons for *d'* between faces, attributes, and pairs, within each set size group were examined. Consequently, the ROC curves show the difference in recognition performance between set size for each test type. Furthermore, the ROC curves consider the relationship between HR and FAR across different confidence thresholds – a variable that was added to the MLM.

The second possible explanation for the results is that the ROC curves attest to the task difficulty. Perhaps there was no difference between the ROC curves because face recognition is a difficult task, irrespective of set size. Figures 4.14 and Figures 4.17 provide support for this explanation. Figure 4.14 shows that recognition performance for faces does not drop drastically across set size, whereas recognition performance for attributes is better than faces - but recognition performance does drop as set size increases and it appears to do so at a greater rate than for faces. In Figure 4.17, it is apparent that the ROC curves for attribute tasks cover a greater area than the ROC curves for faces. Together, the two figures suggest that (a) face recognition performance is worse than attribute recognition performance, and (b) compared to recognition performance for faces, performance for attributes decreases more steadily across set size. Thus, it is not surprising that there are no significant differences between the ROC curves for faces, whereas there was a significant difference between ROC curves for attributes at higher set sizes. It is interesting that there was a significant difference between ROC curves at Set Size 12 and 24 for pairs, especially since both Figure 4.14 and the MLM show that there was no significant difference between faces and pairs at Set Size 12 or Set Size 24. These results suggest that unlike the faces test, the attributes test may become increasingly difficult as set size increases. It is unknown what would happen at higher set sizes: Will attribute recognition performance decrease to the same level as recognition of faces and pairs?

The results of Experiment 1 and Experiment 2 showed differential results for the pairs test. The difference in accuracy for the pairs are most likely due to the type of recognition test used. The HR and FAR for the two experiments are listed in Table 4.24. Inspection of Table 4.24 shows that compared to Experiment 1, HR and FAR are higher and lower, respectively, for Pairs in Experiment 2. This is most likely a consequence of the type of test used in Experiment 2. In Experiment 2, participants' memory for associated pairs was tested with an Old-New task, which could be an easier task than the cued-matching task used in Experiment 1, especially at higher set sizes. For the Old-New task, the base probability of a correct response is 0.5, whereas for the cued-matching task, the base probability of a correct response is 1/Set Size, and this probability decreases as set size increases.

Table 4.24

HR and FAR from Face Recognition Experiment 1 and 2

	Face Recognition Experiment 1						Face Recognition Experiment 2					
	Faces		Attributes		Pairs		Faces		Attributes		Pairs	
SS	HR	FAR	HR	FAR	HR	FAR	HR	FAR	HR	FAR	HR	FAR
1	.94	.07	.99	.01	.99	.01	-	-	-	-	-	-
2	.95	.12	.99	.01	.96	.04	-	-	-	-	-	-
3	.93	.15	.98	.03	.87	.13	-	-	-	-	-	-
4	-	-	-	-	-	-	.72	.21	.93	.03	.70	.12
5	.87	.19	.97	.03	.69	.31	-	-	-	-	-	-
6	-	-	-	-	-	-	.67	.15	.94	.02	.65	.17
8	-	-	-	-	-	-	.60	.14	.95	.05	.52	.20
10	.81	.23	.96	.07	.32	.68	-	-	-	-	-	-
12	-	-	-	-	-	-	.80	.25	.94	.06	.81	.29
15	.76	.26	.95	.05	.27	.73	-	-	-	-	-	-
24	-	-	-	-	-	-	.69	.26	.89	.12	.67	.37
30	.64	.21	.91	.07	.08	.92	-	-	-	-	-	-

Note. HR and FAR indicate Hit Rate and False Alarm Rate respectively. SS indicates Set Size.

The aims of these two experiments were to answer the questions posed at the beginning of this chapter. First, how many faces can humans recognise? Second, how does set size affect face recognition? And third, how does set size affect associative memory (like that between faces and semantic information)?

For the first and second research questions, the reviewed literature suggests that humans are capable of recognising thousands of faces, especially familiar faces, but it is less clear how many unfamiliar faces humans can recognise. From the two experiments presented in this chapter, it appears that face recognition accuracy decreases as set size increases, but not to a great extent. Participants can recognise more than 60% of faces after studying 24 or even 30 perceptually similar faces in one trial. These results are similar to that reported by Metzger (2002): Participants achieved a HR of .67 and FAR of .25 at Set Size 30. It is not clear how participants will perform at set sizes beyond what I have included in this experiment, although there is little doubt that recognition performance will continue to decrease. Additionally, it remains unknown how recognition performance for faces, attributes, and pair will change at set sizes greater than 30: Will recognition performance decrease linearly or nonlinearly, and at what set size will participants perform at chance level (if ever)?

Furthermore, the change in recognition performance as set size increases may reflect more subtly in the criterion levels used by the participants, and the confidence reported by participants. The results from Experiment 2 showed that criterion decreased as Set Size increased, with a criterion level of roughly -0.05 associated with Set Size 30, compared to 0.35, 0.31, and 0.45 for Set Sizes 4, 6, and 8 respectively, and the results from Experiment 1 showed that confidence decreased as set size increased. Together, these two results suggest that (a) participants adopt a more lenient criterion at higher set sizes, and (b) report lower confidence at higher set sizes. These findings provide support for the argument posited by Dobolyi and Dodson (2013) that high criterion will be associated with higher confidence.

Thus, we can conclude that memory performance does decrease as set size increases. The decrease in face recognition performance was more apparent in Experiment 1 than in Experiment 2, although the trend was evident in both experiments. Thus, the present data shows that set size does lead to somewhat reduced face recognition, especially at higher set sizes. These results would need to be replicated, with different stimuli, to determine whether this effect is reliable.

For the third question, we can conclude that associative memory is sensitive to the negative effects of set size, but this is dependent on the type of memory test used. When associative memory was tested with a cued-matching task, recognition memory was significantly worse for pairs than for faces and attributes, respectively, and recognition performance for pairs decreased at a faster rate. When the cued-matching task was replaced by an Old-New test, the significant difference between pairs and faces disappeared, although recognition performance for both tests remained significantly worse than for attributes. The results from Face Recognition Experiment 1 replicate findings that item recognition memory is better than associative recognition memory (Bastin et al., 2010; Bender et al., 2017; Naveh-Benjamin et al., 2004; Naveh-Benjamin et al., 2009), but the results from Face Recognition Experiment 2 only partially replicate these findings, since associative recognition memory was as good as face recognition memory.

The two experiments reported in this chapter contribute in four ways to the literature on associative memory for face-attributes. First, unlike the studies reviewed in this chapter, these two experiments controlled for stimuli characteristics by testing all participants on 30 faces, attributes, and pairs. Regardless of which set size group participants were assigned to, they were required to complete enough trials until they had studied all thirty faces and attributes. This served as a control for stimulus difficulty, and the Set Size 1 group served as a control group as memory was tested under optimal conditions, that is, a lower set size. Secondly, participants were tested on all the items that they encoded, and this method may provide a better estimate of memory performance than testing memory for a subset only. The stimuli characteristics of the subset should be considered when testing recognition. In the current experiment, stimuli characteristics were controlled for by requiring participants to make decisions for each studied item. Third, in both experiments, variation was introduced to the stimuli used at encoding and recognition to prevent picture recognition (Bruce, 1982). The variation between encoding and recognition should have increased task difficulty, since participants would have had less visual information available when making a decision;

however, despite this, participants were still able to accurately recognise faces at fairly good levels (>60% for higher set sizes). Experiments 1 and 2 used the same stimuli, but the poor performance for pairs in Experiment 1 cannot be solely attributed to the variation in stimuli, since Experiment 2 did not replicate the pairs results. Instead, the results are due to the recognition test used in Experiment 2. Finally, this research provides further support that participants perform worse at higher set sizes, as evidenced in ROC curves. This was most pronounced for the pairs test. These results suggest that for higher set sizes, participants are less able to discriminate between Old and New pairs and that their performance shifts closer to chance. Surprisingly, participants show poorer diagnosticity for attributes at Set Size 24 than Set Size 12. These results are attributed to the between set sizes comparisons within test type (which were not included in the MLM), and the overall task difficulty for faces compared to attributes. As speculated in the results section, the recognition test for faces was difficult regardless of set sizes, and therefore there were no differences between set sizes for faces. In comparison, the recognition test for attributes was the easiest of the three tests, and recognition performance was best for attributes. Closer inspection of Figure 4.14 and Figure 4.17 suggest that recognition performance for attributes does decrease as set size increases, especially at higher set sizes (e.g. Set Size 24). These results may suggest that verbal memory is vulnerable to set size, but only at extremely large set sizes.

Limitations

There are some theoretical and practical limitations that are important to consider for this research. First, it was not clear why there were anomalous results for Set Size 12 in Experiment 2: This group appeared to perform better than the immediately previous and subsequent set size groups. I examined the variable data for outliers at both the Hit Rate, False Alarm Rate, and *d'*, but it was not immediately obvious to me who, if anyone, was an outlier.

This result could be a consequence of sampling error, and this anomalous result may have less impact in a larger sample size.

For future research, researchers should replicate the current findings with a new set of faces and attributes. The same stimuli were used in both experiments, and it is not clear whether set size will have the same detrimental effect with different stimuli. For example, future researchers could manipulate stimuli similarity (high similarity, low similarity) to determine whether the set size has the same effect on recognition performance. The faces used in the current experiments were perceptually very similar to each other, and this may have further impaired associative and face recognition memory. It is unknown whether set size would have the same effect on highly dissimilar stimuli.

Furthermore, the sample used in both experiments were heterogenous, as not all participants belonged to the same group as the target images (Malpass & Kravitz, 1969; Sporer, 2001a; Sporer, 2001b). Even though set size demonstrated a significant detrimental effect on face and pairs recognition despite the in-group/out-group identities of the sample compared to the target images, it would be useful to recruit a homogenous group of participants who belong to the same group as the individuals in the target images to better isolate the effect of set size on recognition, without the confounding variable of group-identity. A possible research design could include in-group and out-group conditions where participants and target faces belong to the same and different groups, respectively. Set size may interact with group membership so that recognition performance for faces and pairs are worse for out-group, rather than in-group faces.

Another limitation is that the lineup warning may not have been explicit enough. Participants in the current experiment may not have realised that the target may not be present in the parade, and consequently felt compelled to make an identification. This could have impacted on the lineup accuracy, as well as confidence in the lineup decision. Future research should consider providing a more explicit non-biased lineup instructions.

The final limitation is that ROC analyses require many participants (some estimate roughly 60 per cell; Yonelinas & Parks, 2007). As a result, the ROC curves reported here may be underpowered. For an experiment with five experimental conditions, this would require at least 300 participants.

There are three primary theoretical limitations. First, it was not clear whether the paired recognition test performance was an independent measure of memory. The methods used in the two current experiments were not limitations, but it worth highlighting concerns about the theoretical relationship between the stimuli, and consequently the types of memory tested in the two experiments. The paired recognition test assumed that the participant remembered both face and attribute that comprised that pair; however, if they had not remembered either of the individual components, then they would not be able to respond by choosing the correct attribute from the list or by responding to the correctly paired face and attribute. A possible solution is to test participants only on the faces and attributes that they identified correctly in the two independent item tests; they can be shown their correct choices and asked to pair them with each other. There is, however, a possibility that participants may not correctly recognise any attributes or faces that were paired together originally, resulting in very few, or no stimuli for the matching task. This proposed solution, however, may not adequately address the relationship between associative memory and item memory: Can associative memory exist without independent memories for the items from which it is comprised?

An additional theoretical limitation is that these experiments do not provide the same load at encoding as would be experienced by a participant in an eyewitness experiment. In the current two experiments, participants saw large set sizes of target faces, but each target face was presented only one at a time, whereas participants in an eyewitness experiment would see all the targets committing a crime together. This additional variable – divided attention at encoding – is one of many contributing variables that affect recognition in eyewitness experiments. Recognition performance is worse following divided attention during encoding (Craik, Govoni, Naveh-Benjamin, & Anderson, 1996; Naveh-Benjamin, Craik, Guez, & Dori, 1998), and thus we can hypothesize that a similar pattern of results will be seen in an eyewitness experiment. A comparable face recognition experiment would be one in which all the faces within that set size group are displayed simultaneously at encoding.

` Another limitation is that the eyewitness encoding scenario has a narrative, and this may be more meaningful compared to the face recognition encoding scenario, which is contextless. This narrative may act as a schema by helping eyewitnesses to reconstruct episodic memory for the crime, whereas face recognition provides participants with very few opportunities to develop a cognitive schema to guide recognition performance (Anderson & Pichert, 1978; Arkes & Freedman, 1984; Moscovitch, Cabeza, Winocur, & Nadel, 2016). Consequently, while the current findings suggest that recognition performance is poorer for higher set sizes, this conclusion may not be true for a more meaningful, contextualised scenario, where participants can construct a meaningful narrative of the individual items that comprise the criminal event (e.g. perpetrators, and roles).

Conclusion

In this chapter, I reported on two experiments that contributed to our understanding of how human memory for faces and face-attributes is affected by set size. The results of these studies showed that set size negatively affects face and pairs recognition, but that attribute recognition remained unchanged as set size increased. However, broadly, these results suggest that if presented with a large group of people, like a new class of students, for example, the observer may recognise a randomly selected student, and may recognise a randomly selected name, but is less likely to accurately recall the semantic information associated with the students in that class. This is a common anecdotal experience that most people have experienced. Furthermore, these results suggest that eyewitnesses to multiple-perpetrator crimes may be able to correctly recognise the perpetrator, and may be able to testify to the events of the crime, but may struggle with pairing perpetrators to their respective roles and actions within the crime.

In the next chapter, I will expand this research to investigate how set size affects face recognition and role recollection in an eyewitness experiment. This will address some of the limitations in the current two face recognition experiments, and add the visual complexity of a simulated crime as an encoding task.

Chapter 5

Eyewitness Memory for Multiple Perpetrators

In Chapter 2, I established that eyewitnesses to multiple perpetrator crimes might (a) be expected to make multiple identifications from one or more parades, and (b) are expected to provide supporting information for their identifications but may not be able to do so. In Chapter 4, I investigated how set size affects recognition of unfamiliar faces and face-attribute pairs. The results of the experiments in Chapter 4 demonstrated that memory for faces and attributes were retained independently, although recognition memory for faces was more sensitive to the negative effects of set size. In contrast, the ability to correctly pair attributes and faces was severely hindered by set size: When provided with a cued-matching task, participants were unable to correctly match previously seen attributes with previously seen faces. The results for the pairs recognition task were not replicated when the cued-matching task was replaced with an Old-New task with congruent and incongruent pairs. Instead, results for the pairs recognition task followed the same pattern as results for the face recognition task, and recognition results for both tests were poorer than the recognition results for the attributes test.

In this chapter, I extend the original research question – how does set size impact memory for faces and memory for paired-information – to an experiment that uses an eyewitness paradigm. This links back to the original findings in Chapter 2, and extends the theme of this thesis, which is to investigate eyewitness memory for multiple perpetrators. As mentioned previously, eyewitness experiments and face recognitions aim to answer similar questions about memory but use different methods to do so (Shapiro & Penrod, 1986). Face recognition experiments normally use static images and a contextless encoding scenario, whereas eyewitness experiments use targets who are normally designated the perpetrators and who are embedded within the context of simulated crimes. At test, memory in an eyewitness experiment is tested with an identification parade (where lineup members are presented simultaneously or sequentially), whereas in face recognition research, memory is often tested with an Old-New task or a two Alternative Forced Choice (2AFC) task. The differences in methods between the two research areas can yield different results.

I will begin this chapter with a review of the literature surrounding eyewitness identification in multiple-perpetrator crimes. To date, I have found 15 published manuscripts (12 that use an eyewitness encoding scenario, 1 that investigates eyewitness descriptions of the crime, and 2 that uses images at encoding) and 6 unpublished manuscripts, and I will discuss the overall similarities and differences between these studies. Following this, I will present an eyewitness experiment that investigates eyewitness memory for multiple perpetrators using a lineup recognition task.

What is Known about Eyewitness Memory for Multiple Perpetrators?

Most eyewitness research has investigated eyewitness memory for single-perpetrator crimes. In the typical eyewitness research paradigm, participants view a simulated crime committed by a single individual, whom they need to recognise later. In contrast, there is a dearth of studies investigating eyewitness memory for multiple-perpetrator crimes. It is not clear why this is, especially since researchers of earlier published work commented on multiple-perpetrator scenarios. In 1965, Wall hypothesized that eyewitness memory should be poorer for a crime committed by multiple perpetrators. In fact, he recommended that if the eyewitness states that s/he saw all the perpetrators during the commission of the crime (and made a point of studying them), then any identifications made should be treated with caution. This recommendation appears counterintuitive at face value: surely a conscientious eyewitness who made the effort to study the perpetrators with the aim of recognising them later is a trustworthy, reliable eyewitness? Instead, Wall argued that the encoding of both the crime and the perpetrators cannot be trusted as the eyewitness encoded them with divided attention, and consequently any identification should be dismissed. The only scenario where Wall considers eyewitness identification of multiple perpetrators reliable is one where the eyewitness has ample time to encode each perpetrator.[65] Furthermore, Wells and Turtle (1986) stated in Footnote 2 of their manuscript that the all-suspect parade was one that "does not refer to a situation in which there were multiple offenders. Instead, the situation is one in which there is one offender but multiple suspects. The situation of multiple offenders is not considered in the current article." (Wells & Turtle, 1986, p. 332). Although both sources demonstrate awareness of multiple-perpetrator scenarios, and its consequences for eyewitness memory, there remains a paucity of research on this topic.

Estimator versus System Variables: Definitions and Relationship

The lack of research into eyewitness memory for multiple perpetrators may be due to the perceived importance of system variables in eyewitness research. Wells (1978) first introduced the distinction between system and estimator variables: Estimator variables are characteristics of the crime (e.g., duration), the witness (e.g., intoxication), and the perpetrator (e.g., physical appearance), whose impact on eyewitness memory can only be estimated. System variables form part of the legal and judicial systems, for example, the delay between reporting a crime and giving a statement/interview, building composites, building identification parades; administering identification parades; parade instructions; interviewing methods; and testifying in court. System variables can be manipulated, changed and improved upon to better facilitate eyewitness recall, recollection, and recognition of details of the crime (for example, see Wells et al., 2000); it is, however, not possible - or at the very least, very difficult - to implement measures that change the effect that estimator variables have on eyewitness memory. Wells (1978) argues that the aim of distinguishing between estimator and system

[65] Wall (1986) does not define the optimal encoding length of a multiple-perpetrator crime that will avoid impoverished identification.

variables is not to invalidate research ideas – in fact, he stresses that the utility of the research must be demonstrated by the researcher, and that basic and applied research are necessary. Wells does, however, maintain that applied research that investigates factors known to affect eyewitness memory that are under the control of lawmakers, police officers, and legal stakeholders, have more impact. Cutler, Penrod, and Martens (1987) agree that system variables are important, but also stressed that estimator variables should not be overlooked.

Estimator and system variables, however, may not exist independently of each other. Some estimator variables have an impact on system variables, for example, if the perpetrator rode a bicycle during the commission of the crime then eyewitnesses can be asked to perform a bicycle parade.[66] The tendency to better recognise own-group faces than out-group faces is well established in the literature (Malpass & Kravitz, 1969). Eyewitnesses to crimes committed by out-group perpetrators are more likely to mistakenly identify an innocent suspect from the identification parade. While eyewitness researchers and police officers cannot change the characteristics of the eyewitness and the perpetrator, they can consider the impact of out-group membership while conducting an interview, creating a composite, and building the identification parade. For this reason, I suggest that estimator variables do not exist in isolation. The need for certain police and judicial procedures (i.e., system variables) may arise from the presence of certain estimator variables. Furthermore, police procedure for one level of an estimator variable is not necessarily suitable for other levels of the same estimator, or other estimator variables. Police and judicial procedure should be adapted to account for the impact of different estimator variables on eyewitness memory.

The number of perpetrators who committed the crime is an example of an estimator variable that directly affects various system variables, for example, (a) it directly impacts

[66] Rex v W [1947] 3 All SA 1 (A)

interviewing (which is beyond the scope of this thesis), (b) lineup formation and instructions (refer to Chapter 2, and see Hobson et al. 2012), and (c) and has a unique consequence of role identification, which occurs post-identification. This supports Wells' caution to the reader against adopting a narrow view of system and estimator variables. Instead, researchers should consider how estimator variables influence system variables, and whether changes to system variables are *uniformly* appropriate and beneficial to eyewitnesses of all types of crimes, and are workable solutions that can be implemented in practice. Consider applied research on single-perpetrator crimes, for example. The changes to police procedure that developed from research on single-perpetrator research may not translate into impactful, or even logistically viable outcomes for investigations of multiple-perpetrator crimes. Consequently, the recommended procedures may not yield adequate outcomes (e.g., role assignment, or provide a fair test of eyewitness memory) or police officers do not use the recommended procedures; this was evident in Chapter 2, but will be discussed further in Chapter 7.

Research on Eyewitness Memory for Multiple Perpetrators

To date, I have located 15 published and 3 unpublished manuscripts in the eyewitness literature that investigated eyewitness memory for multiple perpetrators. Of these manuscripts, one is an archival study that investigated physical descriptions of multiple perpetrators (Fahsing, Ask, & Granhag, 2004), two of these studies use images, sans context, at encoding (Bindemann, Sandford, Gillatt, Avetisyan, & Megreya, 2012; Megreya & Burton, 2006) and the remaining manuscripts are experimental studies that have investigated eyewitness memory using a lineup task. The characteristics and results of these eyewitness studies are listed in Table 5.1

Table 5.1

Characteristics and Results of Empirical Research Investigating Eyewitness Memory for Multiple Perpetrator Crimes

Study	Encoding					Recognition						Results	
	Number of Perpetrators	Format of Encoding	Context for encoding?	MA assigned?	Accomplice Roles assigned?	Number of Perpetrators Tested	Parade Format	Suspects / Parade	Foils / Parade	Role Assignment	TA used	HR	FAR
Clifford & Hollin, 1981	1	Video	Yes	Yes	No	1	Photo[a]	1	9	No	No	.35	-
	3	Video	Yes	Yes	No	1	Photo[a]	1	9	No	No	.30	-
	5	Video	Yes	Yes	No	1	Photo[a]	1	9	No	No	.15	-
Dempsey & Pozzulo, 2008	2[b]	Video	Yes	Yes	Yes	2	Photo[a]	1	6	No	Yes	.48	.25[c]
Dempsey & Pozzulo, 2012[d]	2[b]	Video	Yes	Yes	Yes	2	Photo[a]	1	5	No	Yes	.26	.48[c]
Egan et al., 1977	2	Live	Yes	No	No	1	Photo	1	4	No	No	.85	-
						1	Live	1	4	No	No	.98	-
Goldstein, 1977	3	Live	No	No	No	1	Live/ Photo[e]	1	4	No	No	.81	-
Hobson & Wilcock, 2011	3	Video	Yes	No	Yes	3	Photo	1	8	Yes	Yes	.51	.15
Megreya & Bindemann, 2011	1[b]	Video	Yes	Yes	Yes	1	Photo	1	9	No	Yes	.50	.21
	2[b]	Video	Yes	Yes	Yes	1	Photo[a]	1	9	No	Yes	.29	.28
Schiff et al., 1986	6	Video/ photo/ video still	Yes	Yes	Yes	6	Photo	6	11	No	No	.57[f]	-
							Video	6	11	No	No	.66[f]	-
							Other	6	11	No	No	.49[f]	-
Shepherd, 1983	2	Live	Yes	No	No	1	Video	1	8	No	Yes	.20	.42
						2	Video	2	10	No	Yes	.35	
Wells & Pozzulo, 2006	2	Video	Yes	Yes	Yes	2	Photo	1	5	Yes	Yes	.22	.41[c]
Yarmey, 1982	1	Photo	Yes	UK	No	1	Photo	1	20	No	No	.52[g]	-
	3	Photo	Yes	UK	No	3	Photo	3	20	No	No	.14[g]	-
	5	Photo	Yes	UK	No	5	Photo	5	20	No	No	.09[g]	-

Notes. MA = Main Assailant, TA = target absent, HR = Hit Rate, FAR = False Alarm Rate; UK = Unknown. Hit Rates and False Alarms rates are collapsed across over experimental conditions. Unless stated otherwise, FAR is the proportion of suspect identifications from a target-absent parade. The ‘Context for encoding’ column describes whether the encoding materials depicted a specific context (e.g., a crime).

[a]. Participants were told which perpetrator they were meant to identify.

[b]. At least one of the perpetrators was female.

[c]. These studies do not report false alarm rate, but report overall ‘false positives’. These include all foil and suspect identifications.

[d] Dempsey and Pozzulo (2012) tested children’s eyewitness memory.

[e] Results for the two different lineup formats are not reported.

[f] Participants in this could make up to six identifications.

[g] These studies do not report false alarm rate, but report overall ‘false positives’. These include all foil and suspect identifications.

The most frequently cited study investigating eyewitness memory for multiple perpetrators is Clifford and Hollin (1981), but two other studies pre-date this (Egan et al., 1977; Goldstein, 1977). Both earlier studies reported high hit rates – much higher than the rates reported by later studies (see Table 5.1) - and used the smallest parades at recognition (nominal lineup size was five). Goldstein (1977) detailed a study where participants viewed three live targets, and had to identify only one of the targets from a five-person identification parade either two days or two weeks post encoding. Participants did not know that only one target was present in the parade; instead, participants were told that they could make one, two, three, or no identifications. Overall, recognition was extremely high with approximately 81% of participants successfully identifying the target from the parade.[67] False alarm rates were not possible to calculate, because only target-present parades were used; however, Goldstein reported that 44% of participants who successfully identified the perpetrator from the parade also made a second inaccurate lineup decision from the same parade.

In a second experiment, Egan et al. (1977) recruited participants who encoded two live targets for five seconds, but it not clear if the targets were seen consecutively or at the same time. Consecutive encoding may not be completely analogous to simultaneous encoding, and may result in better encoding. Following encoding, participants viewed either a photographic or live parade containing five individuals (one target, four foils). For the recognition task, participants were required to make a decision for each person in the lineup (i.e., they had to answer whether they recognised each person), and they were not specifically instructed that only one suspect would be present in the parade. Again, the parades were always target-present, and like the experiment reported by Goldstein (1977), recognition accuracy was high with 91% of participants correctly identifying the perpetrator. Egan and colleagues reported that 67% of

[67] Goldstein (1977) does not provide the full descriptive statistics for the entire study. This accuracy rate was estimated from the description of the results provided.

participants misidentified a foil from the parade.[68] Only 28% of participants achieved 'complete' accuracy, that is, made a correct identification without a second incorrect response. Although this result may suggest that 28% of participants had a reliable memory of the event, this result could also be interpreted in light of the task and its instructions. The authors stated that when participants asked whether both targets were present in the parade, they were told "You were the only eyewitness and therefore you are the only person who knows" (Egan et al., 1977, pp 202). Consequently, the cohort of participants who achieved flawless recognition accuracy may not have done so because they were most certain that only one perpetrator and that no other perpetrators were present, but instead their recognition performance may have resulted from low criterion response (i.e., not wanting to commit to a second decision and thereby rejecting all other members) due to the ambiguous instructions.

Clifford and Hollin (1981) manipulated the nature of the encoding event and the number of perpetrators between participants. Participants watched either a violent or nonviolent crime that was committed by one, three, or five perpetrators. Afterwards, participants either viewed the identification parade and then completed a questionnaire, or completed the questionnaire before they viewed the identification parade. When viewing the parade, participants were instructed to identify the first perpetrator in the film from a ten-person lineup (nine foils and one target). When completing the questionnaire (referred to as 'testimony' in the manuscript), participants were asked for a physical description of the perpetrator. The authors investigated the effect of encoding context (violent/nonviolent) and number of perpetrators on two dependent measures, testimony accuracy and lineup accuracy. Participants in the nonviolent

[68] Egan and colleagues referred to the error rate as a false alarm rate. This is not a true false alarm rate like that used described in the eyewitness literature. If the parade comprises one target and multiple known-to-be-innocent foils, and the eyewitness identifies the suspect and a foil, there is no reason to assume that one suspect identification and one misidentification will lead the police to investigate the foil. The police normally know that the foils are innocent – in fact, the reason why they are chosen to stand in the parade is because they are known to be innocent so that they provide a true test of the witness' memory (Wells & Turtle, 1986). For this reason, these identifications should rather be treated as errors instead of false alarms.

condition produced significantly better testimony than those in the violent condition, and within the violent condition, participants produced more accurate testimony in the one-perpetrator condition than the ten-perpetrator condition. It was not clear whether the order of the lineup task and testimony had any effect on the dependent variables, but participants who completed the testimony questionnaire after viewing the parade may have committed to their lineup decision, and thus produced better descriptions (if they made a correct decision) or worse descriptions (if they chose the wrong person). This experimental sequence – lineup identification followed by providing testimony – is not used in practice, as eyewitnesses would provide testimony shortly after the crime but before viewing the identification parade.

The finding that eyewitnesses provided worse quality testimony in the violent condition with multiple offenders coupled with the findings that multiple-perpetrator crimes are more violent (Jewkes et al., 2012; Statistics South Africa, 2014) suggest that police officers can expect worse quality testimony and recall of the physical description of the perpetrators. The reader should note that the questions that participants answered in Clifford and Hollin (1981) centred on only the main assailant (whom they had to identify from the parade); for this reason, it is unclear whether this poor testimonial quality would occur for all perpetrators, or if participants recalled less accurate information about the main assailant, because they encoded one of the accomplices instead.

Further identification results from Clifford and Hollin (1981), and other relevant study characteristics are discussed in the relevant sections below.

Control group of a single perpetrator. Three published studies included a single-perpetrator control group in their design (Clifford & Hollin, 1981; Megreya & Bindemann, 2011; Yarmey, 1982), and the control group always outperformed the multiple-perpetrator conditions. Clifford and Hollin (1981) reported that both the single-perpetrator group (mean accuracy = .35), and the three-perpetrator group (mean accuracy = .30) performed significantly better than chance (chance =.10), whereas the five-perpetrator group did not (mean accuracy = .15). Similarly, Megreya and Bindemann (2011) found significant differences between single-perpetrator and two-perpetrator conditions: Accuracy in the single-perpetrator conditions was lower than in the two-perpetrator conditions (.29 versus .54 respectively). Yarmey (1982) did not report any inferential statistics comparing the hit rate between the three set size conditions (one, three, and five perpetrators); the results, however, suggested that participants in the single-perpetrator condition (.52) performed better than participants in the three-perpetrator condition (.14), who performed only marginally better than the five-perpetrator condition (.09).

Of the three studies that included a one-perpetrator control group, only Megreya and Bindemann (2011) incorporated target-absent (TA) parades. Their results showed that the detrimental effect of multiple perpetrators at encoding was limited to the hit rate, since they found no significant difference in the false alarm rate between the single-perpetrator and two-perpetrator conditions. Overall, these three studies suggest that compared to a single-perpetrator control group, eyewitnesses perform worse at a recognition task following encoding of a multiple-perpetrator crime.

Number of perpetrators at encoding and recognition. Across the studies listed in Table 5.1, the number of multiple perpetrators presented at encoding ranged between two (Dempsey & Pozzulo, 2008; Dempsey & Pozzulo, 2012; Egan et al., 1977; Megreya & Bindemann, 2011; Shepherd, 1983; Wells & Puzllo, 2006) and six (Schiff et al., 1986). For most of the studies that tested two perpetrators, hit rate ranged between .20 (Shepherd, 1983) and .48 (Dempsey & Pozzulo, 2008), except for research reported by Egan et al. (1977), whose participants achieved an accuracy rate of between .85 and .98. Their results are anomalous when compared to the hit rate reported in other experiments - in fact, Egan et al. (1977) reported a hit rate that was at least double that reported elsewhere. Such a high hit rate is even more unexpected given that Egan and colleagues manipulated the delay between encoding and recognition so that it varied between 2, 21, and 56 days. These results, however, should be interpreted alongside the rest of the reported lineup results. Egan and colleagues required that their participants make a decision for every parade member, and from this they tallied the ‘false alarm’ rate. They use the term ‘false alarm’ rate to denote any mistaken identification of another known-to-be-innocent lineup member (i.e. a foil) rather than reserving this term for identification of an innocent suspect from a target-absent parade. Of their participants, only 28% did not misidentify any other foils, but 67% made at least one mistaken identification. The ‘false alarm’ rate increased from 48% to 62% to 93% across a delay of 2, 21, and 56 days respectively. Such high results for both HR and FAR suggest that participants may have adopted a lenient criterion– participants were, after all, given ambiguous instructions about the number of suspects who were present in the parade, and from the manuscript it does not appear that participants were warned that the suspect/s may not be present. Additionally, little information is given about the composition of the live parade and photographic parade used, except that the foils and targets resembled one another - however it is not clear to what extent. Another explanation for the anomalous results could be participants may have experienced reduced divided attention if the perpetrators were presented consecutively rather than simultaneous, and consequently, participants had a better encoding opportunity and subsequently better recognition. Megreya and Burton (2006) present counter-research to this hypothesis and demonstrate that encoding is superior for faces presented simultaneously than sequentially (see page 199 of the current manuscript).

Of the three experiments that used more than five perpetrators at encoding, HR was .15 (Clifford & Hollin, 1981) and .09 (Yarmey, 1982) for five perpetrators, and between .49 and .66 for six perpetrators (Schiff et al., 1986). Neither of these studies incorporated a TA condition. Of the three studies, Clifford and Hollin tested recognition of only one perpetrator from the total set used at encoding, whereas the remaining two studies tested recognition for all the perpetrators shown at encoding by placing all the perpetrators in a single parade with either 11 foils (Schiff et al., 1986) or 20 foils (Yarmey, 1982) The HRs reported by Yarmey (1982) were conservative, because it reflected only participants who identified all the targets. In contrast, Schiff and colleagues reported the overall HR for any perpetrator – on average participants achieved a recognition rate of 4.07 of the six perpetrators (this translates into the 68% [4.07/6] for the HR reported in the manuscript). Without a one-suspect control group, however, it is unclear whether the encoding and recognition stimuli used by Schiff et al. (1986) were too easy. In contrast, both Clifford & Hollin (1981) and Yarmey (1982) used two additional set size groups (one-perpetrator and three-perpetrator) and their results showed that HR decreased as set size increased (see Table 5.1).

Recognition of all perpetrators. Of the experiments listed in Table 5.1, six tested recognition memory for all the perpetrators shown at encoding (Dempsey & Pozzulo, 2008; Dempsey & Pozzulo, 2012; Hobson & Wilcock, 2011; Schiff et al., 1986; Shepherd, 1983; Wells & Pozzulo, 2006; Yarmey, 1982). The results of two of these studies must be interpreted with caution. First, Dempsey and Pozzulo (2012) tested children's recognition memory of multiple perpetrators, whereas the other six experiments involved adult participants. The results reported by Dempsey and Pozzulo may not generalise to adult eyewitnesses, since child eyewitnesses behave differently to adult eyewitness (for example, see Lindsay, Pozzulo, Craig, Lee, & Corber, 1997; Pozzulo & Warren, 2003). Second, unlike the remaining studies, Hobson and Wilcock (2011) manipulated lineup format within participants. Their participants watched a video of a simulated crime committed by three perpetrators, and completed a recognition task afterwards. During the recognition task, their participants always saw a combination of TP and TA parades, but never saw only TP or TA parades. Consequently, it is not possible to estimate how many participants recognised all the perpetrators. Hobson and Wilcock did, however, report that only 11.11% of participants made three correct decisions across the three parades.

Overall, across the six studies, few participants could accurately recognise all perpetrators: Shepherd (1983) reported that only one of forty-one participants (2.4%) correctly identified both perpetrators - a much lower estimate than the HR (.35) listed in Table 5.1, which represents the proportion of participants who made at least one correct identification. Similarly, the HR listed in the table for Wells and Pozzulo (2006) is the overall accuracy rate, averaged across both lineup parades, and does not reflect participants who made only correct decisions. Instead, Wells and Pozzulo reported on page 423 that only 10.67% of participants could correctly identify both targets. Yarmey (1982) reported that 14% and 9% of participants could identify all the perpetrators in the three-perpetrator and five-perpetrator conditions, respectively. Schiff et al. (1986) did not report how many participants identified all six perpetrators.

Encoding context, role assignment, and memory for roles. Of the published studies listed in Table 5.1., only one did not embed the encoding scenario within a criminal context by portraying the target individuals as perpetrators of a simulated crime (Goldstein, 1977). Of the remaining studies, six designated one target as the main assailant (Clifford & Hollin, 1981; Dempsey & Pozzulo, 2008; Dempsey & Pozzulo, 2012; Megreya & Bindemann, 2011; Schiff et al., 1986; Wells & Pozzulo, 2006), and of these, all except one (Clifford & Hollin, 1981) assigned roles to the accomplices. Only one study assigned roles to all the perpetrators without designating any one perpetrator as the main assailant (Hobson & Wilcock, 2011).

Of those studies that assigned roles to all the targets, two tested for role assignment (Hobson & Wilcock, 2011; Wells & Pozzulo, 2006). Wells and Pozzulo (2006) showed their participants a video of a two-perpetrator crime where the perpetrators were designated as assailant or accomplice. In the video, the assailant provides instructions to the accomplice; afterwards, a female victim appears and while the accomplice distracts her, the assailant grabs her bag and both perpetrators flee. Following a 25-minute distractor task (where the participants provide a written description of the event), participants were shown one of three lineup parade types: either two simultaneous parades, two sequential parades, or a new lineup variant termed the two-person serial lineup. For the two-person serial lineup, participants were shown pairs of photographs (either one of the two targets and a foil, or only foils); the photographs within each pair was shown together, one pair at time. Participants were shown either TP or TA parades, but never both. Their task was to identify the two culprits and then state what role the culprits performed in the video. Wells and Pozzulo (2006) reported that role assignment was not problematic – in fact, only one of the 33 participants, who correctly identified at least one of the targets, was unable to recall the role of the identified target.

Hobson and Wilcock (2011) showed participants a video of a simulated crime committed by three perpetrators, and each perpetrator was assigned a unique role or attribute: one drove the vehicle, one carried the rucksack, and the third one, who was seated in the backseat of the getaway vehicle, carried a laptop bag. During the recognition stage of their experiment, half of the participants were asked to reflect on the role of the perpetrator while viewing the parade,[69] and the remaining participants were given standard lineup instructions[70] with no mention of role. Participants made their lineup identifications on a response sheet, and if they made a positive identification, also identified the role of that target. Of the two groups of participants, roughly 69% of participants who considered the role while viewing the parade correctly recalled the role of perpetrator, whereas 30.3% of participants who received general instructions correctly recalled the roles. Role instructions had no effect on lineup accuracy. The authors do not provide any further interpretation of these results, and it is not clear why participants performed better at role recollection when considering the roles while viewing the

[69] Participants were told to “identify what role that perpetrator played in the crime event and mark their identification decision on the response sheet and complete all the questions” (Hobson & Wilcock, 2011, p. 291).

[70] Participants were told to “mark their identification decision on the response sheet and complete all the questions” (Hobson & Wilcock, 2011, p. 291).

parade versus when only considering the role after an identification is made. One reason for the improved role recollection is that participants who received role instructions adopted a more lenient criterion to make a role identification, and consequently, there was a higher proportion of correct responses. These explanations are speculative since the descriptive statistics for role identification were not reported.

Other Relevant Studies: Unpublished Manuscripts and Matching Tasks

There are eight other studies that are relevant for discussion of eyewitness memory for multiple perpetrators. Two of these are published manuscripts that used matching tasks (Bindemann et al., 2012; Megreya & Burton, 2006), and six are unpublished student theses (Hobson, 2013; Jacob, 1994; Laldin, 1997; Owen, 2009; Tupper, 2018; Vanderwal, 1996).

Three of the earlier unpublished student theses were produced by RCL Lindsay's research laboratory at Queen's University[71] and used the same materials (Jacob, 1994; Laldin, 1997; Vanderwal, 1996). At encoding, participants watched a 45-second video of three perpetrators stealing a woman's purse. Each study manipulated the lineup format used at recognition. Jacob (1994) gave participants three six-person simultaneous parades or one sequentially presented lineup comprising 24 individuals; either one, two, three, or none of the perpetrators were present in the parade/s. Vanderwal (1996) also manipulated simultaneous and sequential parades between participants. In the simultaneous parade condition, participants were shown a single parade comprising 18 individuals. In the sequential condition, participants were shown three sequential parades, one for each target, comprising five, six and seven pictures respectively. For both parade conditions, either one, two, three, or none of the perpetrators appeared in the parade/s. In the final experiment, Laldin (1997) utilised an unusual variant of the sequential and simultaneous parades, which was similar to the two-person serial lineup variation employed by Wells and Pozzulo (2006): Instead of showing participants

[71] http://www.queensu.ca/psychology/People/Faculty/Roderick-Lindsay.html

individual images, with one person in each image, Laldin (1997) grouped the lineup members into triads, and each triad acted as one lineup option. Therefore, participants viewed multiple triad images, and made a decision for each triad. If participants responded affirmatively to a triad, they were subsequently asked to indicate which persons (one, two, or all three) in the triad were the targets. Both TP and TA parades were constructed, and within the TP conditions, either one, two, or three perpetrators appeared together in the same triad. Laldin further manipulated lineup presentation so that half of the participants were assigned to either a simultaneous or sequential lineup condition. In the simultaneous lineup condition, participants viewed six triads presented simultaneously, and in the sequential lineup condition, photographs of each triad were presented one at a time. Furthermore, participants were warned that if one or more perpetrators were present then they would appear together in the same triad. The results of the three studies are shown in Table 5.2.

Table 5.2

Hit Rate and False Alarm Rate Reported in Jacob (1994), Laldin (1997) and Vanderwal (1996)

		Hit Rate			False Alarm Rate		
Study	Parade Format	1 Perpetrator Present	2 Perpetrators Present	3 Perpetrators Present	0 Perpetrators Present	1 Perpetrator Present	2 Perpetrators Present
Jacob, 1994	Simultaneous	.50	.39	.30	.11	.11	.11
	Sequential	.33	.25	.30	.11	.08	.06
	Average	.42	.32	.30	.11	.10	.09
Laldin, 1997	Simultaneous	.43	.64	.44	.17	.26	.24
	Sequential	.24	.31	.24	.0	.0	.05
	Average	.34	.48	.34	.09	.13	.14
Vanderwal, 1996	Simultaneous	.05	.20	.31	.07	.07	.07
	Sequential	.08	.22	.14	.53	.13	.13
	Average	.07	.21	.23	.30	.10	.10

Note. Participants saw parades where either one, two, three, or no perpetrators were present. Thus, Hit Rate and False Alarm Rate are calculated from the recognition task for each participant, and depending on the study, even from the same parade. Participants made a false alarm when they mistakenly identified an innocent suspect from the parade. For this reason, the False Alarm Rate was only calculated for the 0, 1, or 2 Perpetrators Present condition (there were no innocent suspects in the 3 Perpetrators Present condition). For the same reason, Hits were not calculated for the 0 Perpetrators Present condition.

The simultaneous parade presented in Vanderwal (1996) is most similar to the parades used by SAPS officers (as a discussed in Chapter 2) - SAPS officers reported that they would often place all suspects in one parade together. When compared to the simultaneous parade used by Jacob (1994) where each suspect was placed in their own parade, the trend suggests that (a) overall, participants made more hits when one suspect, rather than multiple suspects, was placed in a single parade (.40 versus .19); (b) participants were better able to identify one suspect in a smaller parade than a larger parade (.50 versus .05), (c) but the FAR was higher when each parade contained only one suspect rather than multiple suspects (.11 versus .07).

Three other unpublished student theses were produced from three different laboratories (Hobson, 2013; Owens, 2009; Tupper, 2018). Owens (2009) reports a similar design as Jacob (1994), Laldin (1997), and Vanderwal (1996): Participants viewed a crime committed by one, two, or three perpetrators, and after providing a statement about the crime and a short filler task, viewed either TP or TA, simultaneous or sequential parades. Overall, Owens reports that lineup accuracy decreased as the number of perpetrators increased (one-perpetrator condition: 10%; two-perpetrators condition: 6%; three-perpetrators condition: 3%). These results are in line with the results reported in Table 5.1. Compared to sequential parades, Owens found that simultaneous parades were more diagnostic for main assailant when the main assailant was present in the parade. These interesting results coupled with the results listed in Table 5.2. suggest that future research should investigate the effect of parade format of identification of multiple perpetrators. Hobson (2013) produced an interesting study that manipulated the number of perpetrators who committed the crime (1 versus 3), as well as the number of parades seen at recognition (1 for single-perpetrator, and multiple-perpetrator versus 3 for multiple-perpetrator). Thus, Owens was able to compare the effect of the names at encoding between the single-perpetrator and multiple-perpetrator groups, and was able to compare the effect of the number of lineup decisions that had to be made within the multiple-perpetrator groups.

Overall, her results suggest that participants in the single-perpetrator conditions performed better than either of the multiple-perpetrator conditions in the TP conditions, but there was no difference between the recognition performance of groups in the TA condition. Interestingly, there was an effect of load at encoding when load at recognition was controlled: Participants in the single-perpetrator condition who viewed one TP parade made more accurate decisions than participants in the multiple-perpetrator condition who viewed only one TP parade. There was no significant difference in lineup accuracy for TA parades. Hobson reports similar results to published studies that calculates accuracy for all perpetrators (Dempsey & Pozzulo, 2012; Hobson & Wilcock, 2011; Schiff et al., 1986; Shepherd, 1983; Wells & Pozzulo, 2006; Yarmey, 1982): None of the participants in the three-perpetrator condition were able to accurately recognise all three perpetrators. In a second study, Hobson further investigates the mechanism underlying the poor facial recognition for multiple-perpetrator crimes by employing a change-blindness experiment. Compared to a single-perpetrator condition, participants in the multiple-perpetrator condition were less likely to recognise when one of the actors was replaced by another actor during encoding (10% versus 55%). These results suggest that the poor recognition performance following multiple-perpetrator crimes is due to divided attention. Tupper (2018) extended the multiple-perpetrator identification question to determine whether later showup decisions for a multiple-perpetrator crime were contingent on previous showup decisions. Unlike lineups where the witness views an array of individuals of which only one is the suspect, showups consist of only one individual who may be the suspect. Like Hobson (2013), significant results were only found for the target-present parades. Specifically, participants were more likely to make a decision for the third showup if (a) that showup was TP, and (b) they had made a decision on the previous parade. Unfortunately, these results were not replicated in Experiment 2: Participants were more likely to make a decision if the third showup was TP, regardless of whether they made a decision for showup 2.

Up to this point, the reviewed literature tested eyewitness *memory* of a multiple-perpetrator event using an eyewitness memory paradigm. Both Bindemann et al. (2012), and Megreya and Burton (2006) provide evidence that the detrimental effect of encoding multiple faces is not limited to recognition performance, but also evident in matching performance. Megreya and Burton report four experiments that followed roughly the same format: Participants completed multiple trials where they saw either one or two target faces, which they had to recognise from a 10-person array. Their results showed that recognition performance was worse following the presentation of two faces even when participants were given unlimited time to study the faces (Experiment 1), and that this two-face disadvantage persisted when the recognition task was replaced with a matching task (Experiment 3) – this result suggests that the mere presence of a second face is enough to impair performance, perhaps by increasing attentional load.

Bindemann et al. (2012) replicated the two-face disadvantage with matching and recognition tasks using the same materials as those used by Megreya and Burton (2006). In experiment one of both manuscripts, hits, misses, and misidentifications differed significantly between the one-target and two-target conditions, with participants performing worse on all three measures in the two-target conditions. Furthermore, Bindemann et al. tested whether the two-face disadvantage was a consequence of participants not knowing *which* of the two target faces would appear in the parade. In one of their experimental conditions, participants viewed two faces, one outlined in red (i.e., the foil) and the other outlined in green (i.e., the target), and then performed the matching task. They found that following the cued encoding, recognition performance for the TP and TA parades fell between that observed in the one-target and two-target conditions in the matching task in Experiment One (Bindemann et al., 2012). The only significant difference was between misidentifications reported for the one-target and two-target cued condition: Participants made more misidentifications even when the target image was outlined and performed similarly (albeit slightly better) to the two-target conditions where target and foil were not outlined. When the matching task was replaced with a recognition task, participants in the two-target cued group yielded more hits than participants in the two-target uncued group (61.3% versus 48% respectively) and did not differ from the one-target group. The authors posit that these results suggest that the two-face disadvantage arises from a combination of perceptual and memory demands.

In the limitations section of Chapter 4, I posited that the sequential presentation of the faces at encoding was a possible limitation, especially since eyewitnesses to a crime were unlikely to view each perpetrator one at a time. Megreya and Burton present evidence that lineup performance is, indeed, affected by the presentation mode of multiple target images. In Experiment 2, Megreya and Burton (2006) manipulated whether the two target images were presented sequentially or simultaneously prior to the delayed matching task (i.e., the task was presented two seconds after encoding). Overall, their results suggest that performance for the delayed matching task was superior when targets were encoded simultaneously rather than sequentially: Hits were significantly lower for faces presented sequentially rather than simultaneously (31.2% versus 42%), and false alarms were higher for sequentially than simultaneously presented faces (39.2% versus 26%). Megreya and Burton’s findings suggest that recognition performance was superior for faces presented simultaneously, rather than sequentially, at encoding; this contradicts the limitation that I offered in Chapter 4. These findings challenge the divided-attention argument, but continue to demonstrate the detrimental effect of set size at encoding. The authors go on, however, to state that reason for the poor performance is poorly understood, especially since participants were given unlimited time to encode the faces.

Aim and Rationale

From the reviewed eyewitness memory literature, it appears that hit rate decreases as set size increases. However, only three studies included a single-perpetrator control group (Clifford & Hollin, 1981; Megreya & Bindemann, 2011; Yarmey, 1982), and of these, only one incorporated a target-absent condition (Megreya & Bindemann, 2011). Although Megreya and Bindemann did report that FAR was unaffected, this result has not been replicated.

From the literature, only seven published manuscripts tested recognition for all the perpetrators seen at encoding (Dempsey & Pozzulo, 2008; Dempsey & Pozzulo, 2012; Hobson & Wilcock, 2011; Schiff et al., 1986; Shepherd, 1983; Wells & Pozzulo, 2006; Yarmey, 1982); however not all studies report recognition rates for all perpetrators, and instead report accuracy rates for only one perpetrator. Recognition rates may be inflated for conditions where accuracy is reported for at least one accurate identification. From the few studies that did report the proportion of participants who could accurately recognise all the perpetrators, it appears achieving 100% accuracy is difficult, as very few participants were able to do so.

Only two studies tested for role recall and assignment (Recall: Hobson & Wilcock, 2011; Recall and assignment: Wells & Pozzulo, 2006), and in both studies, participants could correctly pair roles with perpetrators. Neither of these studies manipulated set size. Hobson and Wilcock tested role recall for three perpetrators, and Wells and Pozzulo tested recall and role assignment for two perpetrators. The findings from Chapter 4 showed that pairing accuracy declined steeply after Set Size 5. Thus, it remains unclear whether role pairing decreases as the number of perpetrators increases, as we saw in the face recognition experiments reported in Chapter 4.

The aims of the current experiment arise from the current review and from the findings in Chapter 4. First, this experiment aims to determine how set size affects recognition ability of eyewitnesses. The face recognition experiments reported in Chapter 4 suggest that there is a tipping point at Set Size 5 with decreasing recognition ability from that point onwards; for this reason, the current experiment will include multiple set size groups that increase up to ten perpetrators. Second, the current experiment will include a one-perpetrator control to serve as a comparison group to the multiple-perpetrator crime conditions. It is hypothesised that recognition performance will be worse for the multiple-perpetrator conditions than the single-perpetrator conditions, and that within the multiple-perpetrator conditions, recognition will decrease as set size increases. Finally, this experiment aims to determine how set size affects role recollection. Thus, when participants make an identification from the lineup, they must also recall the role performed by that individual.

Method

Design

This experiment had two factors. Set size was a between-subjects factor with five levels, and described the number of perpetrators (one, two, three, five, or ten) who committed the crime. There was a second factor, lineup condition, with two levels: Condition A and Condition B. For participants in the one-perpetrator Set Size group, lineup condition was a between-subjects factor: Half of the participants saw a target-present parade (Condition A), and the remaining half viewed a target-absent parade (Condition B). For all participants who encoded a multiple-perpetrator crime, lineup condition (target-present, target-absent) was manipulated within-subject. For example, for the two-perpetrator Set Size group, half of participants saw Perpetrator 1 – TP, and Perpetrator 2 – TA, and the other half saw Perpetrator 1 – TA, and Perpetrator 2 – TP. These lineup conditions are illustrated in Table 5.3.

Table 5.3

Assignment of Condition A and Condition B of Lineup Format Across Set Size

		Lineup Format for Lineup Condition A									
Set Size	n	Perpetrator 1	Perpetrator 2	Perpetrator 3	Perpetrator 4	Perpetrator 5	Perpetrator 6	Perpetrator 7	Perpetrator 8	Perpetrator 9	Perpetrator 10
1	20	TP	-	-	-	-	-	-	-	-	-
2	20	TP	TA	-	-	-	-	-	-	-	-
3	19	TP	TA	TP	-	-	-	-	-	-	-
5	21	TP	TA	TP	TA	TP	-	-	-	-	-
10	20	TP	TA	TP	TA	TP	TA	TP	TA	TP	TA

		Lineup Format for Lineup Condition B									
Set Size	n	Perpetrator 1	Perpetrator 2	Perpetrator 3	Perpetrator 4	Perpetrator 5	Perpetrator 6	Perpetrator 7	Perpetrator 8	Perpetrator 9	Perpetrator 10
1	20	TA	-	-	-	-	-	-	-	-	-
2	20	TA	TP	-	-	-	-	-	-	-	-
3	21	TA	TP	TA	-	-	-	-	-	-	-
5	19	TA	TP	TA	TP	TA	-	-	-	-	-
10	20	TA	TP	TA	TP	TA	TP	TA	TP	TA	TP

Note. TP and TA denote target-present and target-absent parades respectively.

Sample

Participants (n = 200) were recruited using the SRPP (Student Research Participation Programme) system at UCT. Their average age was 20.02 years (SD = 2.28), and the sample comprised 56 women.

Materials

Encoding videos. Ten white South African men (all between 18 and 35 years) depicted the perpetrators in a simulated crime. This crime was a theft of computer equipment from a university computer laboratory. In the one-perpetrator condition, which served as the foundation for the other videos, a man is seen peering through the glass door at the entrance of the laboratory. He opens the door, enters the laboratory, and checks whether the other doors in the laboratory are unlocked. Meanwhile, the camera has a close-up view of his face for three seconds. Afterwards he turns around to approach a computer cubicle, where he removes his backpack, places a computer keyboard inside, picks up a computer monitor and rushes out of the computer with the stolen items. Altogether, the entire video was 15 seconds long. (see Appendix L for a video-still from the one-perpetrator video).

Additional versions of the simulated crime were filmed where the main assailant was accompanied by either one, two, four, or nine accomplices (comprising levels two, three, five and ten of factor, set size). Each accomplice was assigned a unique role that they adopted across all films (see Table 5.4). To prevent attentional divide among perpetrators within a set time of 15 seconds and to allow eyewitnesses a reasonable opportunity to encode each perpetrator, each version of the crime was lengthened by 15 seconds for each additional perpetrator.[72]

[72] It is extremely difficult to control for stimuli difficulty and encoding opportunity. If each version of the film was 15 seconds long – regardless of the number of perpetrators – then recognition performance is confounded by set size and divided attention. Allotting 15 seconds to each additional perpetrator aimed to reduce attentional load; however, this introduced the possibility that the stimuli material was not standardised across conditions and participants were afforded more opportunity to encode the perpetrators in the multiple-perpetrator conditions.

Table 5.4.

A List of Unique Roles Performed in the Various Set Size Crime Videos

	Set size				
Roles performed by perpetrators	1	2	3	5	10
Main assailant stole the computer screen	x	x	x	x	x
Accomplice 1 rummaged through the drawers of a filing cabinet		x	x	x	x
Accomplice 2 filmed the crime with his cellular phone			x	x	x
Accomplice 3 drew on the walls with chalk				x	x
Accomplice 4 acted as a lookout by the door				x	x
Accomplice 5 rolled and smoked a cigarette					x
Accomplice 6 sat atop a small bar fridge and drank from a can of Coca-Cola					x
Accomplice 7 threw toilet paper across the laboratory					x
Accomplice 8 rolled the chairs across the laboratory					x
Accomplice 9 used a set of keys to try unlock all the doors and cabinets					x

Note. Set size refers to the number of perpetrators who commit the crime together. The symbol 'x' denotes that the perpetrator was present in that condition.

In the multiple-perpetrator versions of the crime, each perpetrator approached the camera, one at a time, so that the viewer had a close-up, frontal-view of the perpetrator for roughly three seconds (all materials can be found online).

In total, 10 films were recorded (two versions of each of the five set size conditions). In the two versions of the films of each set size condition, different actors portrayed different perpetrators. This was an attempt to introduce variability across the two versions of the set size condition, and to avoid confounding perpetrator with actor identity.

Each film was filmed in high definition, and in colour. The camera was in the same position for each film, and the length of the films varied from 15 to 150 seconds for the one-perpetrator and ten-perpetrators conditions, respectively. The films were shown without audio to the participants.

Lineups. Target-absent (TA) and target-present (TP) lineups were constructed for each target using a modified version of the procedure recommended by Malpass, Tredoux and McQuiston-Surrett (2007). Ten mock-witnesses briefly viewed each target face for three seconds, and then, following a short delay, provided a written physical description of that face. Each mock witness provided a description of each target. Together with a research assistant, we created modal descriptions for each target from these descriptions, and used these modal descriptions and the photographs of the targets to independently choose 10 possible foils for each target from a database of photographs curated by Tredoux. After making our foil selections, we compared our selections, and foils that we had both chosen were placed in that target's parade. No foils appeared in more than one parade. The target-present parade contained six photographs: five foils and one target; whereas the target-absent parade for that same target contained the same five foils, but the target was replaced with a suspect who was randomly selected from that target's original foil selection. Additionally, two versions of each TP and TA parade were constructed where target position was changed between versions. Lineup position was randomly chosen for each version.

In total, 40 lineups (10 targets x two lineup formats [TP, TA] x two position versions) were constructed; each lineup contained six individuals. Each lineup was in colour, and the photographs depicted frontal-view, neutral expression faces of the targets and foils. Pictorial artefacts were digitally removed from each photograph, and photographs were edited so that a white t-shirt was superimposed at the neckline, and target image was placed in front of a grey, stucco background (see Figure 5.1 for an example, and see online link for all the identification parades).

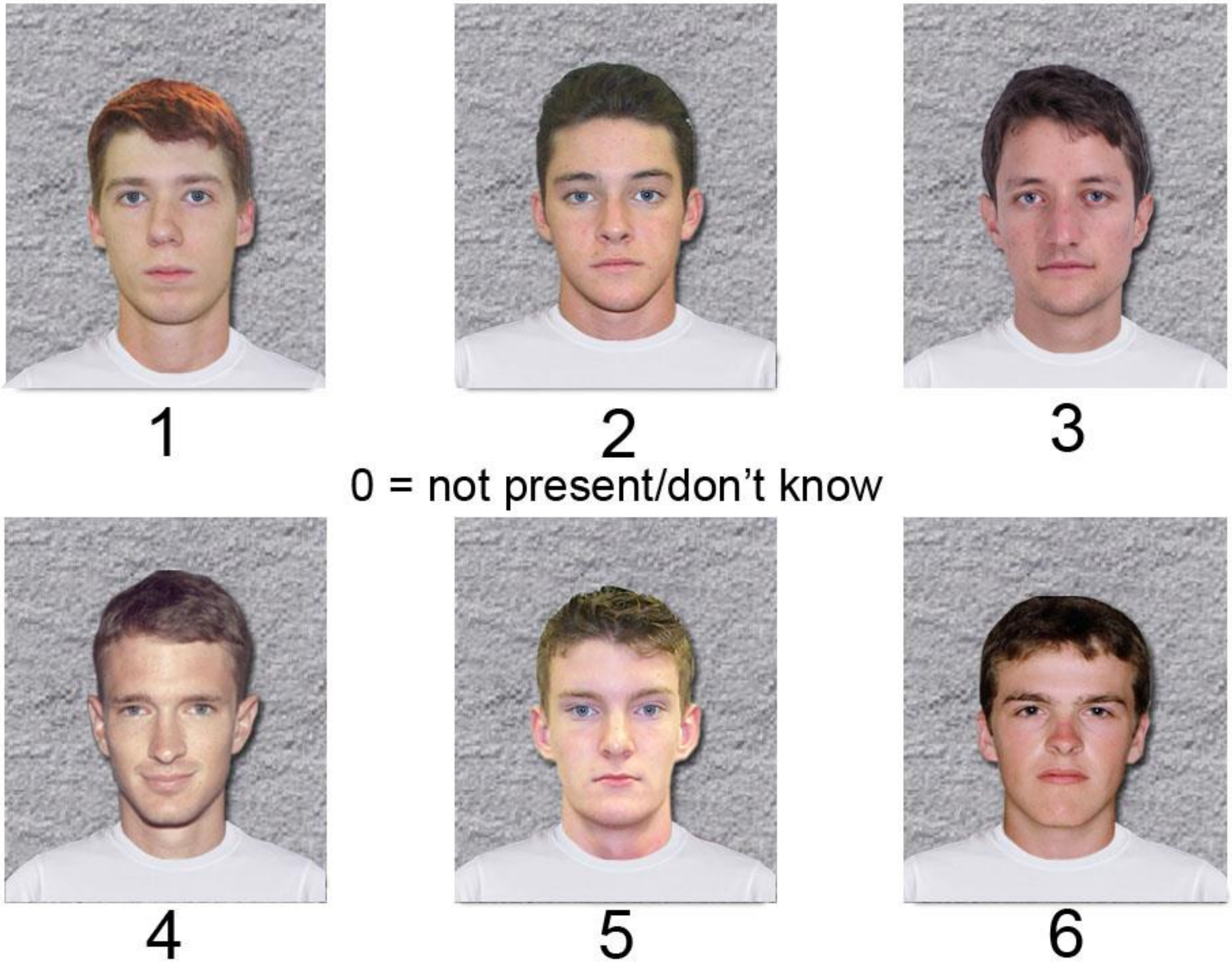

Figure 5.1 Target-present version of Target 9, who was the main assailant in version B of all the videos. The target is in position 3.

Procedure

Encoding

Participants were tested in groups of one to six individuals. After signing the consent form, each participant was seated at their own computer. The experiment was administered using E-Prime 2.0.8 (Psychology Software Tools, Pittsburgh, PA), and all responses were collected electronically using a keyboard. Participants were told that they would watch a video of a crime that took place in one of the computer labs on campus. When the video finished, participants completed a filler task for five minutes. [73]

Statement

After completing the filler task, participants were told to assume that they were reporting the crime to the police station and had to provide a written statement about the 'crime' in the film. Following this, participants in the one-perpetrator and multiple-perpetrators conditions were asked five or six questions about the crime respectively (the sixth question asked if any of the perpetrators appeared to be in charge; see Appendix M). Afterwards, participants completed a second distractor task, which took roughly 30 minutes.[74]

Recognition

The recognition phase began after the second distractor task. Participants in both the one-perpetrator and multiple-perpetrator conditions were given similar lineup instructions and were warned that the perpetrator may or may not be present (see Appendix N for the instructions), and then viewed the lineups. For each lineup, participants had unlimited time to respond. Participants made their lineup decision by pressing the number key that corresponded to the number that appeared below the face or pressed '0' for 'not present'. After making any

[73] During this filler task, participants wrote an essay defining intelligence accompanied by examples of intelligent behaviour in humans.

[74] During the second filler task, participants were instructed to document all examples of intelligent behaviour displayed by the octopus in a nature video (see the video here: https://www.youtube.com/watch?v=lBLRCs5Xobg). These filler tasks were chosen so that participants were not exposed to any additional faces between encoding and recognition.

lineup decision, participants were asked to rate their confidence in their decision on a scale from 0 to 10, where 0 indicated lowest confidence and 10 indicated highest confidence.

If participants made a positive identification, i.e., they identified someone from the lineup, they were asked to state the role that person performed in the crime. All participants were asked this, even those in Set Size 1. Typed responses were collected digitally via a message box. Following role recollection, participants were asked to rate their confidence in their role recollection on a scale from 0 to 10. Participants who made a positive identification were given metacognitive questions to probe their decision-making process. Participants who rejected the identification parades skipped the role recollection phase and were given a different set of metacognitive questions to investigate how they made their decision. Confidence and these questions were not analysed in this thesis.

For participants in the one-perpetrator condition, the recognition phase ended when they answered the metacognitive questions; however, participants in the multiple-perpetrator conditions repeated the recognition trial (from lineup instructions to metacognitive questions) until they had viewed all the lineups for their experimental condition. All participants saw the identification parade for the main assailant first so that recognition accuracy for the main assailant could be compared across conditions while controlling for lineup sequencing effects and response biases (much like the procedure used in Podd, 1990). Subsequent lineups were shown in a randomised order.

Once the experiment was completed, participants in the multiple-perpetrator condition were given a questionnaire about how they experienced the lineup tasks. These questions were not analysed in this thesis.

Results

Lineup responses for TP parades were scored as Hits, Misses, and Incorrect Rejections, and for TA parades scored as False Alarms and Correct Rejections. The total number of foil and suspect misidentifications from the TA parade were divided by the nominal lineup size (6) to compute the False Alarm Rate (FAR). This method was used because a suspect was not purposefully chosen based on how well they matched the written description or physical appearance of the target (see Meissner et al., 2005).

Lineup Accuracy for the Main Assailant

Lineup responses for the main assailant are listed in Table 5.5. Inspection of this table shows that the average HR for the main assailant is higher in the single-perpetrator condition (.55) than the multiple-perpetrator conditions (.35). A z-test for two independent proportions showed that the difference in HR was marginally significant, $Z = 1.64$, $p = .051$. Thus, HR for the main assailant was lower for the multiple-perpetrator conditions than the single-perpetrator condition. The average FAR is higher in the multiple-perpetrator conditions (.11) than in the single-perpetrator condition (.08). Additionally, *d'* is higher in the single-perpetrator condition compared to the multiple-perpetrator conditions (1.53 versus 0.87), and *c* is lower in the multiple-perpetrator conditions, indicating that participants adopt a more lenient criterion (i.e., they are more willing to make a decision).[75]

Table 5.5

Hit Rate and False Alarm Rate for the Main Assailant

	TP			TA			
SS	Hits	Misses	Incorrect Rejections	Correct Rejections	Total FA (/6)	*d'*	*c*
1	.55 (11)	.25 (5)	.20 (4)	.50 (10)	.08 (10)	1.53	-0.64
2	.33 (7)	.48 (10)	.19 (4)	.32 (6)	.11 (13)	0.79	-0.83
3	.40 (8)	.40 (8)	.20 (4)	.50 (10)	.08 (10)	1.15	-0.83
5	.37 (7)	.26 (5)	.37 (7)	.43 (9)	.10 (12)	0.95	-0.81
10	.30 (6)	.55 (11)	.15 (3)	.20 (4)	.13 (16)	0.60	-0.83
MP	.35	.42	.23	.36	.11	0.87	-0.82

Notes. TP and TA denote target-present and target-absent conditions. FA stands for False Alarms. MP implies multiple perpetrator, and this row shows the average values for all the multiple-perpetrator groups. The *d'* and *c* reported here is calculated from the average hit rate and false alarm rate at the group level. Thus, there is no variance. FA is a weighted false alarm where the total number of suspect and foils identifications is divided by the nominal lineup size.

[75] The signal detection values *c* and *d'* reported in this chapter may be inflated by the weighted FAR. By dividing the FAR by 6, the FAR is quite reduced and there is very little change. This will affect the values of *d'* and *c* for this chapter, since both equations use the FAR in their formulae. I am not certain what is the best method to correct this, but eyewitness researchers should be aware of this when calculating signal detection methods from weighted FAR.

Results were analysed with logistic regression, where set size predicted accuracy. Accuracy for TP and TA parades were analysed separately. Neither logistic regression model was significant, χ^2 (4) = 3.08, p = .55 and χ^2 (4) = 4.93, p = .29 for TP and TA parades respectively. Thus, there was no significant difference in HR and FAR for the main assailant across set size.

Lineup Accuracy and Role Accuracy for All Assailants

The descriptive results for all assailants across the Set Sizes are listed in Table 5.6. Participants saw as many parades as perpetrators, and these parades were a combination of TP and TA parades (except for Set Size 1 where participants viewed only one parade, which was either TP or TA). No participants in the multiple-perpetrator conditions saw only TA or TP parades. For Set Size 3, half of the participants viewed a combination of one TP parade and two TA parades, and the other half viewed two TP parades and one TA parade. Similarly, participants in Set Size 5 each saw an unequal number of TP and TA parades. For this reason, the rates for the different lineup responses were calculated participant-wise, and then averaged by set size group.

Closer inspection of the descriptive statistics for the TP parades show that HR appears to decrease, whereas Missed and Incorrect Rejections increase, as set size increases (see Table 5.6). For the TA parades, Correct Rejections appear to increase, whereas False Alarms remain roughly constant, as set size increases.

Table 5.6

Descriptive Statistics for All Lineup Responses Across Target-Present, and Target-Absent Parades.

	Target-Present			Target-Absent		Signal Detection Measures	
Set Size	Hits	Misses	Incorrect Rejections	Correct Rejections	False Alarms	*d'*	*c*
1	.55 (0.08)	.25 (0.07)	.20 (0.06)	.25 (0.07)	.08 (.10)	1.53 (NA)	0.64 (NA)
2	.38 (0.08)	.48 (0.08)	.15 (0.06)	.38 (0.08)	.10 (.01)	0.61 (0.08)	0.30 (0.04)
3	.43 (0.07)	.34 (0.07)	.24 (0.06)	.44 (0.07)	.09 (.01)	0.92 (0.08)	0.40 (0.05)
5	.29 (0.05)	.41 (0.06)	.30 (0.05)	.44 (0.06)	.10 (.01)	0.80(0.10)	0.71 (0.05)
10	.28 (0.04)	.38 (0.04)	.35 (0.04)	.45 (0.04)	.09 (.01)	0.74 (0.11)	0.97 (0.05)

Notes. Standard errors are in parentheses. False alarm rate is calculated from the sum of all TA identifications divided by 6. There are no standard errors reported for *d'* and *c* for Set Size 1, because these values were calculated from the group means for Hits and False Alarms.

Comparing perfect accuracy for lineups across set sizes. In the studies reviewed in the literature, very few tested eyewitness memory for all perpetrators; however, of those that did, most reported poor recognition performance (ranging between 2% and 22%). To judge how participants performed in this experiment, overall accuracy was computed for each participant. Overall accuracy included all Hits and Correct Rejections. Within Set Size 1, 52.5% of participants achieved perfect accuracy; within Set Sizes 2, 3, and 5 the number of participants who achieved perfect accuracy decreased to 15%, 17.5%, and 2.5% respectively (see Table 5.7.). No one in Set Size 10 achieved perfect accuracy; the highest overall accuracy achieved was eight (out of ten), and only one participant achieved this score.

Binomial tests showed that participants in all Set Size groups performed better than chance, all *ps* < .05.

Table 5.7

Proportion of Participants who made Only Correct Lineup Decisions

SS	Highest Score Achieved	n	Proportion	CI Lower	CI Upper	Chance
1	1	21	.525	.370	.680	.143
2	2	6	.150	.039	.261	.020
3	3	7	.175	.057	.293	.003
5	5	1	.025	0.00	.073	<.0001
10	8	1	.025	0.00	.073	<.0001

Note. SS denotes Set Size. CI stands for Confidence Intervals. The highest score possible is equal to the set size. For Set Size 10, the probability of obtaining a score of 8 is .000006.

Proportions between consecutive set sizes groups were tested with binomial tests. Binomial tests showed that participants in Set Size 1 performed better than Set Size 2, $p = .006$. Remaining binomial tests showed no significant differences in proportions between Set Sizes 2 and 3, 3 and 5, and 5 and 10; *p*s > .05.

Overall lineup accuracy. Two one-way ANOVAs were calculated for *d'* and *c* for the multiple-perpetrator groups (Set Size 2, 3, 5, and 10) respectively. It was not possible to include Set Size 1, because there was no variance in *d'* and *c*, which were calculated from the means for the between-subject TP and TA conditions. For both *d'* and *c*, homogeneity of variance was preserved. There was no significant main effect of set size on *d'*, $F(3, 156) = 2.14$, $p = .100$.

There was, however, a significant main effect of set size on *c*, $F(3,156) = 40.45$, $p < .001$. The group means were plotted with confidence intervals to allow for further investigation (see Figure 5.2). There was no significant difference in *c* between Set Size 2 and 3, $t(69.829) = -1.34$, $p = .18$. The difference in criterion between Set Sizes 3 and 5 was significant, $t\,(77.657) = -4.334$, $p < .001$, and between Set Sizes 5 and 10, $t(77.989) = -3.7043$, $p = .001$. Participants in Set Size 10 adopted a stricter criterion than participants in Set Size 5, who adopted a stricter criterion than participants in Set Size 3.

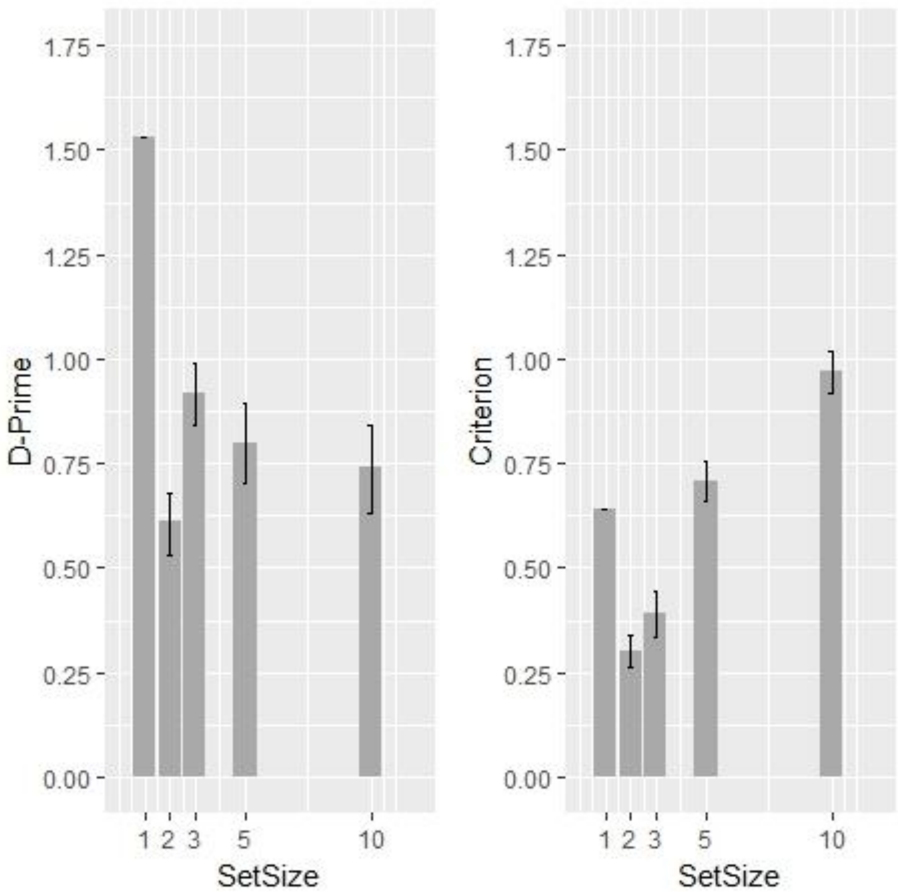

Figure 5.2. Average *d'* and *c* for participants across Set Size groups. The error bars are 95% confidence intervals. There are no error bars for Set Size 1 because these signal detection measures were calculated at the group level from the group average for Hits and False alarms that were measured between subjects.

Accuracy for overall lineup identification across all five set size groups. There was no significant difference between *d'* for the multiple-perpetrator conditions (Set Size 2 – 10), but this did not include a comparison with the single-perpetrator condition. To compare lineup performance across all levels of Set Size, *d'* was reduced to its individual components, Hits and False Alarms.

First, a mixed effects logistic regression was used to analyse the accuracy (i.e., Hits) for TP parades where accuracy was nested within participant.[76] Two models were built: A null model where accuracy was predicted by the intercept and random effects for participant, and a second model where accuracy was predicted by both the random effects for participant, and the fixed effects for set size. The difference in chi-square between the two models was significant, χ^2 (4) =10.12, *p* = .039. The model coefficients listed in Table 5.8.

Table 5.8

Coefficients for Fixed Effects and Random Effects for Model 2 Predicting Accuracy for TP Parades

Fixed Effects	Estimate	CI lower	CI upper	Odds Ratio	CI lower	CI upper	p
Intercept	0.231	-0.783	1.245	1.26	0.457	3.478	.655
Set size 2	-0.819	-2.076	0.438	0.44	0.125	1.551	.202
Set size 3	-0.438	-1.626	0.751	0.65	0.197	2.119	.471
Set Size 5	-1.227	-2.391	-0.063	0.29	0.091	0.939	**.039**
Set Size 10	-1.332	-2.443	-0.221	0.26	0.087	0.802	**.019**
Random Effects							
Participants							
Number of observations	419						
Number of participants	180						
ICC	0.175						
Variance	4329.514						

Notes. The reference group is Set Size 1. Emboldened text indicates p < .05

[76] There are varying data points nested within Participant for each level of Set Size. Therefore, for Set Size 1, there is only 1 data point per Participant, but up to five for Participants in Set Size 10. This could mean that the betas are not reliable for the lower set size groups. However, when Set Size 1 is removed, and the mixed effect logistic regression is rerun, the variance for the random effect remains the same, suggesting that the Set Size 1 has little influence on the random effect.

When accuracy for TP parades was predicted by Set Size, the probability of obtaining a Hit was significantly lower at Set Size 5 than Set Size 1, *OR* = 0.29, *z* =-2.066, *p* = .039, and at Set Size 10 than Set Size 1, *OR* = 0.26, *z* = -2.350, *p* = .019. The probability of making a correct decision in Set Size 1 was 4.35 times higher than in Set Size 5, and 4.85 times higher than in Set Size 10. These results may appear anomalous compared to Table 5.6, but Table 5.6 compares the overall hit rate, whereas Table 5.8 lists the logit of making a correct or incorrect decision for TP parades. Furthermore, the percentage reported in Table 5.6 does not reflect the repeated decisions that each participant made, nor does the proportional value of each decision have the same value between set size groups. This will be discussed further in the limitations of the chapter (see Appendix O). Furthermore, consecutive contrasts showed no significant differences, all *p*s > .05. All p values were adjusted for multiple comparisons.

A second mixed effects logistic regression was used to predict false alarms for TA parades. This model was not significant, χ^2 (4) = 1.1106, *p* = 0.89, and not investigated further.

Accuracy for Overall Pairing Identification Across All Five Set Size Groups

I coded all the role responses with the assistance of a second coder using a coding scheme that we developed. For this task, we coded whether the role provided by the participant was a role that was performed within the video. There was strong agreement between us, κ = .849, 95% CI [.824,.877], *p* < .001. After this, we coded whether the participant had made a correct pair, that is, whether the correct role was provided for the correct perpetrator. There was strong agreement on the coded responses, κ = .810, 95% CI [.777, .841], *p* < .001.

To analyse pairing accuracy, the data was manipulated in the following way: Only Lineup Hits were selected for each set size group, and then from the Lineup Hits, I calculated the proportion of responses where the role was correctly recalled correctly identified targets. Thus, the proportion of correct pairs is dependent on the number of Hits and is only possible for TP parades (see Figure 5.3).

Inspection of Figure 5.3 suggests that pairing accuracy decreases as Set Size increases. Within Set Size 1, all the participants who achieved a Hit also correctly recalled the role of the perpetrator. This is expected. Within Set Sizes 2 and 3, of the participants who achieved Hits, 73.3% and 78.6% also successfully recalled the roles respectively. For Set Sizes 5 and 10, however, only 53.3% and 49.1% of participants who made correct identifications also correctly recalled the roles respectively.

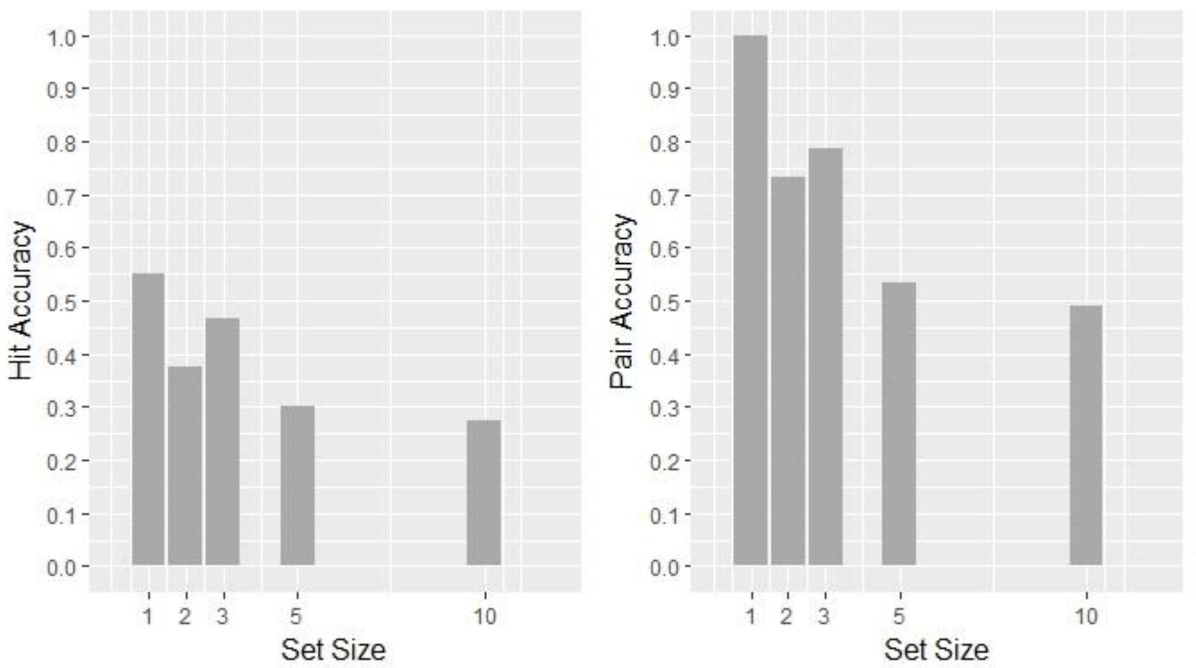

Figure 5.3. Accuracy for lineup identifications and pairing. The image on the left shows the average Hit Rate for TP parades across Set Size, and the image on the right shows the average accuracy for role recollection for the Hits reported in the left image. Confidence intervals are not reported for these figures (see Appendix O).

On average, the participants in the multiple-perpetrator groups performed worse at the role pairing (63.58%) than participants in the Set Size 1 group (100.00%). A test of two independent proportions showed that the difference in proportions was significant, p = .019. I wanted to analyse the pairing accuracy across the five set size groups with a mixed effects logistic regression, but participants in Set Size 1 achieved 100% pairing accuracy (see Table

5.9); thus it was not possible to fit this model as Set Size 1 was overfitted. For this reason, I removed Set Size 1 from the analyses.

Table 5.9

Frequencies of Participants Who Achieved Either Correct or Incorrect Pairing

	Correct Pairing		Incorrect Pairing	
Set Size	N	%	N	%
1	11	100.00	--	
2	11	73.33	4	26.67
3	22	78.57	6	21.43
5	16	53.33	14	46.67
10	27	49.09	28	50.91
MP average	76	63.58	52	36.42

Note. Percentages are calculated row-wise. Within Set Size 1, no participants made any incorrect pairing decisions. Data reported here is partitioned within participants who achieved a hit. Pairing was correct if the participant also achieved a hit.

A mixed effects logistic regression was used to compare pairing accuracy across the four multiple-perpetrator set size groups. Two models were constructed: A null model where pairing accuracy was predicted by the random effects for participant and a second model where pairing accuracy was predicted by the random effects for participant and the fixed effects for set size. The second model was a significantly better fit than the null model, $\chi^2(3) = 8.45$, $p = .038$. The coefficients of the second model are shown in Table 5.10.

Table 5.10

Coefficients for Fixed Effects and Random Effects for Model 2 Predicting Pairing Accuracy for TP Parades

Fixed Effects	Estimate	CI lower	CI upper	Odds Ratio	CI lower	CI upper	p
Intercept	1.063	-0.184	2.311	2.90	0.83	10.08	.095
Set size 3	0.280	-1.245	1.806	1.32	0.29	6.08	.719
Set Size 5	-0.923	-2.379	0.533	0.40	0.09	1.70	.214
Set Size 10	-1.127	-2.558	0.305	0.32	0.08	1.36	.123
Random Effects							
Participants							
Number of observations	128						
Number of participants	84						
ICC	0.065						
Variance	0.230						

Notes. CI denotes Confidence Intervals. The reference group is Set Size 2.

Post-hoc contrasts were run to further investigate whether there were any significant differences between Set Sizes 2 and 3, 3 and 5, and 5 and 10. Consecutive contrasts were chosen due to the ordinal nature of the Set Size groups. None of the consecutive contrasts were significant, *p*s > .05.

Discussion

Most of the studies reviewed in the multiple-perpetrator eyewitness literature only tested recognition for the main perpetrator (Clifford & Hollin, 1981; Megreya & Bindemann, 2011) or for only one of the perpetrators shown at encoding (Egan et al., 1977; Goldstein, 1977; see Table 5.1). Clifford and Hollin (1981) found that participants in Set Size 1 and 3 performed better than chance, whereas participants in Set Size 5 did not. Two of the studies reported very high accuracy rates, with more than 80% of participants correctly identifying the perpetrator from the TP parade (Egan et al., 1977; Goldstein, 1977). Megreya and Bindemann found a significant difference in the HR and FAR between their single-perpetrator and double-perpetrator conditions, in the direction expected: Performance was worse (i.e. lower HR, and higher FAR) in the double-perpetrator condition.

To isolate the effect of set size on lineup recognition for the main assailant, I employed a similar method as the one used by Podd (1990): The first parade shown to all participants was the TP or TA parade for the main assailant. The difference between the overall HR for the single-perpetrator and multiple-perpetrators groups was significant, whereas there was no difference in overall FAR for the single-perpetrator and multiple-perpetrator groups. Logistic regression showed that there was no difference in HR or FAR across the five set size groups. From the group averages of HR and FAR, *d’* was calculated, and results suggested that *d’* decreased as Set Size increased. It was not possible to test this (since there was no variability within *d’* for each group), but *d’* does suggest that overall performance decreased due to the additive changes within Hits and False Alarms, which are not evident when these measures are examined in isolation.

Seven of the reviewed studies tested recognition memory for all perpetrators (Dempsey & Pozzulo, 2008; Dempsey & Pozzulo, 2012; Hobson & Wilcock, 2011; Schiff et al., 1986; Shepherd, 1983; Wells & Pozzulo, 2006; Yarmey, 1982). The results of these studies are important, because they inform how well eyewitnesses would perform at recognising all the perpetrators in a multiple-perpetrator condition. Of these, Hobson and Wilcock (2011) reported that 11% of their participants achieved perfect accuracy (on a combination of TP and TA parades) following the encoding of three perpetrators. For only TP parades, Shepherd (1983), and Wells and Pozzulo (2006) reported perfect accuracy results of 2.4% and 10.6%, respectively, following encoding of two perpetrators.

In the current experiment, fewer participants achieved perfect accuracy in the multiple-perpetrator (approximately 8.75%) than in the single-perpetrator conditions (52.5%). Within the two-perpetrator condition, 15% of participants achieved perfect accuracy,

Participants in the two-perpetrator condition of the current experiment performed marginally better than participants in Shepherd (1983; 2.4%) and Wells and Pozzulo (2006; 10.6%). There are multiple reasons for the discrepancy between the results of the three studies. For example, the encoding and recognition materials in the current study may have been easier than in Shepherd, or Wells and Pozzulo. The three studies also used different parade formats. Participants in the Set Size 2 condition of the current study always saw a TP and a TA parade, whereas both comparison studies used two TP parades (Shepherd, 1983; Wells & Pozzulo, 2006). It may be more difficult to correctly identify two perpetrators from two TP parades, than to make one accurate decision for a TP parade and a TA parade, since participants can reject the TA parade accurately for many reasons and not only because they definitively know that the perpetrator is not present.

For Set Size 3 in the current study, roughly 17.5% of participants obtained perfect accuracy; this result does not differ much from the 11% reported by Hobson and Wilcock (2011). From Size 5 onwards, however, very few participants achieved perfect accuracy: Only one participant (2.5%) of those in Set Size 5 achieved complete accuracy, and no one in Set Size 10 achieved the highest or second highest score. Overall, these results suggest that eyewitnesses to multiple-perpetrator crimes may struggle to correctly recognise all the perpetrators from the identification parades, especially if the crime included five or more perpetrators. Consequently, rather than asking eyewitnesses to identify all perpetrators, it may be better if police only test eyewitnesses on the perpetrator whom the eyewitness is most confident.

The results of the current experiment provide counter-evidence for the suggestion made by Wall (1965) that eyewitnesses who view longer multiple-perpetrator crimes are more reliable than eyewitnesses who view multiple-perpetrator crimes of a shorter duration. Even though I did not manipulate the length of the crime in this experiment within Set Size groups, I did control for the amount of time allotted to each of the multiple-perpetrator conditions. Despite the increased exposure time at encoding, participants in Set Size 5 and 10 were less likely to accurately identify the perpetrator than participants in Set Size 1. Furthermore, previous research suggests that the detrimental recognition performance for multiple faces persists even when participants are given unlimited time to study the face (Megreya & Burton, 2006) or when they are told which face to study (Bindemann et al., 2012). Together, these results suggest that the recognition disadvantage associated with multiple perpetrators could be a consequence of poor encoding or perceptual load, which results in poor memory performance.

Only three of the studies reviewed included a control group where participants encoded only one perpetrator (Clifford & Hollin, 1981; Megreya & Bindemann, 2011; Yarmey, 1982). For all three studies, descriptive statistics showed that recognition memory was poorer for the multiple-perpetrator conditions than the single-perpetrator condition. The inclusion of a one-perpetrator control group is vital. First, it is necessary to demonstrate the double-perpetrator disadvantage (Megreya & Bindemann, 2011), where participants perform worse at recognition tasks following encoding of two perpetrators, compared to a one-perpetrator condition. Second, it acts as a control of the encoding and recognition stimuli – if participants perform better in the multiple-perpetrator conditions than the single-perpetrator conditions, then this suggests an error with the instructions or stimuli. For the current experiment, a single-perpetrator condition was included as a comparison group for the multiple-perpetrator groups. Descriptive statistics showed that fewer participants achieved Hits in TP parades as set size increased, whereas False Alarms remained largely unchanged. Results of the inferential statistics showed that participants were less likely to obtain Hits in Set Sizes 5 and 10 compared to the single-perpetrator condition, even when the perpetrator was present in the lineup. There were no significant differences between set sizes in the probability of making a correct rejection for TA parades. There was also a significant effect of set size on the criterion statistic among the multiple-perpetrator conditions. These results suggest that participants who had viewed more perpetrators were more likely to adopt a stricter criterion, and were more likely to reject the parade.

Only two of the studies reviewed tested role-assignment (Hobson & Wilcock, 2011; Wells & Pozzulo, 2006). Both studies reported that participants did not struggle to recall the roles of the perpetrators – in fact, Wells and Pozzulo (2006) reported that only one participant was unable to recall the role performed by the identified perpetrator. In the current experiment, approximately 73.33% and 78.57% of participants in the two-perpetrator and three-perpetrator conditions, respectively, could correctly recall the role of the perpetrator whom they had identified. In Set Size 5, approximately 53.3% of participants who made a Hit could also recall the role accurately, and only 49.09% in the Set Size 10 group could correctly recall the role.

Overall, fewer participants in the multiple-perpetrator group, than in the single-perpetrator group, were able to correctly recall the role of the perpetrator; however, all post-hoc tests showed no significant differences between multiple-perpetrator groups.

Limitations

The aim of the current study was to better understand eyewitness memory for multiple perpetrator crimes, especially that of lineup identification and role pairing identification. The current experiment developed from the results of the face recognition experiments reported in Chapter 4, which demonstrated that item memory was better preserved than associative memory, and this was especially evident when a cued matching task was used for the pairing task. The design of the current eyewitness experiment differed from that used in the face recognition experiments. Role recollection responses were dependent on the lineup task: Participants were first asked to make a lineup identification, and if they identified someone, they were asked to recall the role of that person. This is a different pairing task to that used in the face recognition experiments, but it is more analogous to the task used by the police – it is highly unlikely that the police will give the witness their statement and ask them to match the role from their statement to the perpetrator that they identified. The method of the current experiment has two possible limitations. First, it is not clear how an eyewitness makes an identification. For example, eyewitnesses may only recognise perpetrators whose roles they recollect from the crime, as evidenced in Hobson and Wilcock (2011): Participants who were instructed to think of the role of the perpetrator while viewing the parade performed better at the recognition task than participants who did not receive these instructions. When coding the roles, the coding assistant and I noticed many repeated roles across the larger set sizes, suggesting that participants were guessing what role that perpetrator performed. Second, the data points for the two tasks (lineup recognition, role recollection) were not independent of each other, and this complicated the data analysis. The pairing data was partitioned from the

Hits data, subsequently reducing the number of possible data points for analysis. For future studies, researchers should make sure to include large sample sizes to ensure that there is enough power.

Another limitation is that I did not compare face recognition (i.e., lineup performance), role recognition, and pairing recognition as in the previous two face recognition experiments reported in Chapter 4. In the face recognition experiments, face recognition accuracy and pairing accuracy were collected from two independent tests. It was not possible, however, to compare lineup accuracy and pairing accuracy with each other in the current experiment, since the pairing data was partitioned from the lineup accuracy. Furthermore, it is not clear how to isolate role recognition/recollection in this experiment: I could analyse the statements provided, but these results would not be similar to the face recognition experiments (because the statements are a verbal recollection of an entire event instead of an Old-New recognition test of attributes), nor the lineup task (because the statement takes place after a much shorter delay than the lineup task following the crime). One could reframe the results from the Face Recognition Experiments in Chapter 4 as evidence that participants have stronger memory for items than associative memory, and this was replicated in the both face recognition experiments. In contrast, the results from the current eyewitness experiment suggest that at the lineup task, participants are *less* likely to correctly identify all the perpetrators as Set Size increases, *less* likely to correctly identify any perpetrators in TP parades, and *less* likely to correctly recall their roles when they do recognise them. What remains unanswered is the reason for this result: Why are participants in multiple-perpetrator set size groups less likely to accurately recall the role of the recognised perpetrator? This question will be addressed in Chapter 6.

There are other limitations that should be addressed in future research. For example, only simultaneous parades were used in this parade, and each parade contained only one perpetrator. The results from Chapter 2 show that police officers often use larger parades that contain at least two, but often all of the suspects. It remains unanswered whether the results from the current study will be replicated in a research design where perpetrators appear in the same parade. The results from the meta-analysis by Shapiro and Penrod (1986) show a significant relationship between Load at Recognition and FAR, but this was in the opposite direction expected, that is, Load at Recognition decreased FAR. The reason for the reduced FAR is unclear: Are more participants rejecting the parade because they are confused or because they know that the perpetrator is not present? If, however, Load at Recognition did have a beneficial relationship with FAR, then police officers would benefit from using larger parades where more suspects are present. Vanderwal (1996) showed participants one large parade containing multiple suspects, and they did report lower FAR than Jacob (1994) who showed multiple simultaneous parades, each containing one suspect. However, lineup format for the two simultaneous parades was not manipulated within the same experiment, and thus it is not clear if there are any other confounding variables. This suggestion is speculative, and needs to be further tested in the laboratory.

There are other research questions that this study was not able to answer. It was very difficult to control the characteristics of the materials in this study. The length of each video was increased by 15 seconds for each additional perpetrator to control for the confounding effect that divided attention has on recognition ability. Consequently, recognition performance for the current experiment may be overestimated: None of the eyewitness studies reviewed in this chapter controlled for attention and encoding across set size groups, and this may explain the low recognition results reported in the reviewed literature. A second consequence of increasing the video length was that the crime and the roles therein became more complex, and this may explain the poor role pairing performance. Thus, encoding complexity may be a confounding variable, but the very nature of multiple-perpetrator crimes is that these crimes are more complex – the effect is not limited to only the number of perpetrators, but includes the interactions among perpetrators and the complex narrative that arises from this.

The need to control stimulus characteristics extends to the materials used at recognition. I used an adapted version of the guidelines outlines provided by Malpass, Tredoux, and McQuiston-Surrett (2007) - these guidelines are suited for constructing materials for single-perpetrator research scenarios and do not address all the difficulties that arise from multiple-perpetrator research scenarios. For example, the descriptions generated from the mock witness task were not unique enough to avoid the same foils being chosen for different targets, and for this reason, I decided to build the parades using the descriptions and photograph of the target. Additionally, to build 10 parades with six physically similar foils, who are also unique to each parade, requires access to a large enough sample of photographs.

The descriptive results from the current experiment showed that participants in Set Size 3, compared to Set Size 2, achieved a higher overall Hit Rate (.43 versus .38), higher d' (0.92 versus 0.61), higher total accuracy (.175 versus .150), and slightly higher pairing accuracy (78.57% versus 73.33%). This result was surprising. A possible reason could be stimulus distinctiveness: One of the targets in the three-perpetrator condition was taller than the other targets, and for this reason, was more distinctive. Although the actors for the current experiment were matched on certain characteristics (e.g., young adult men, under the age of 35 years old, who identified as 'white' and were South African), other perceptual characteristics were not controlled (e.g., distinctiveness, attractiveness, criminality, or baby-facedness). Future research should consider manipulating the target characteristics to determine whether the detrimental effect of set size is negated by target distinctiveness. Research by Megreya and Bindemann (2011) suggests that face recognition is impaired following a two-face encoding scenario even if the two targets are perceptually distinct from each other – in their experiment, participants encoded a pair of perpetrators who were male and female, and the female perpetrator wore a head scarf (which should increase the perceptual disparity between the two targets). The differences between Set Size 2 and Set Size 3 found in this experiment attest to the importance of measuring and controlling stimuli characteristics.

A final consideration for future research is to increase the delay between encoding and recognition. In the current experiment, the delay was 30 minutes, which is quite short and is unlikely to resemble a real-world scenario. Multiple-perpetrators crimes are complex, and it is unlikely that the police may find all the suspects and hold the parade within 30 minutes. It is difficult to estimate what effect delay might have on recognition performance, especially for role pairings, although I suspect that recognition performance would be reduced; however, a more realistic delay of, for example, 24 hours could be more informative of eyewitness memory.

Conclusion

This experiment aimed to test the effect of set size on eyewitness memory by (a) examining lineup recognition performance for all the perpetrators who committed the crime together, and (b) by testing recollection of role pairing. The results from the two face recognition studies in Chapter 4 suggested that memory for individual items was better preserved than associative memory, which was very sensitive to the negative effects of set size. The results from the current eyewitness experiment showed that lineup recognition was worse for larger set sizes, and that participants were not able to correctly recall the roles of all the correctly identified perpetrators. All participants in Set Size 1 could recall the role; however, the odd ratios of correctly recalling the role dropped from 2.90 in Set Size 2 to 1.32, 0.40, 0.32 in Size 2, 3, 5 and 10, respectively.

These results have important consequences for police investigation: They suggest that eyewitnesses of multiple-perpetrator crimes will struggle to correctly identify all the perpetrators from the lineup task. These results also suggest that eyewitnesses to multiple-perpetrator crimes may not be able to recall the role performed by that perpetrator – *even if* they correctly identified a perpetrator from the parade. For this reason, the police and courts should place less emphasis on the eyewitness' ability to correctly recall the role, and recognise that the eyewitness identification may still be correct despite role confusion. This is a controversial point, because the aim of testing eyewitness memory is to demonstrate that eyewitnesses do, indeed, have that memory (and are not guessing), that their identifications and testimony are not incorrect, and to provide the courts with an estimate of the strength of their memory. Knowing that eyewitnesses to multiple-perpetrator crimes may not be able to accurately recall the role of the identified perpetrator means that less weight should be given to eyewitness testimony about the role identification, and provides support for the judicial solution of using the principle of common purpose for sentencing in multiple-perpetrator crimes (as described in Chapter 1).

Chapter 6

A Revised IAC Model for Face and Person Recognition of Multiple Perpetrators

In previous chapters, I investigated face recognition memory for multiple targets using two face recognition experiments and an eyewitness experiment. For all three experiments, the results showed that face recognition decreased as set size increased. Furthermore, the results also demonstrated that associative memory was impaired by an increase in set size. In this chapter, I provide a computational explanation for the impaired recognition performance for faces and associative memory by revising the interactive activation and competition (IAC) model reported by Burton et al. (1990). Burton et al. used the IAC model to account for semantic priming with faces, and I have adapted the model to provide a computational explanation for the findings of my research. In the current chapter, I will review the relevant models of face recognition, introduce the IAC model, and explain how I adapted the IAC model for eyewitness recognition of multiple-perpetrator crimes.

Models of Face Recognition

The extant literature investigating face recognition is substantial. More than thirty years ago, Shapiro and Penrod (1986) included 128 manuscripts in their meta-analysis, but the number of available published manuscripts has since increased exponentially: A Google Scholar search using the terms ‘face recognition eyewitness memory’ yields approximately 4 000 results. A number of findings from face recognition and eyewitness experiments have informed public policy (e.g., Wells et al., 2000) and advanced scientific knowledge about how faces are perceived and remembered. Fewer manuscripts have posited an explanation of the findings through models of face recognition.

Four types of face recognition models. The existing models that aim to explain how humans recognise faces can be categorised into four categories (Rakover & Cahlon, 2001): Functional models, multidimensional models, models that use principal component analysis

(PCA), and connectionist models.[77] All four types of models aim to explain face recognition but use a different approach. Functional models provide theoretical explanations without any computational simulations (Bruce & Young, 1986; Hay & Young, 1982 as cited in Hay, Young, & Ellis, 1991). The face recognition model proposed by Bruce and Young (1986) is the most well-known and will be discussed in more detail later in this chapter.

Multidimensional models aim to explain how the different perceptual dimensions of faces interact with each other to represent where in the vector space[78] the face exists. For example, Valentine (1991) proposed a multidimensional space with the following dimensions: distinctiveness, inversion, and race. Within the multidimensional space, faces – which are represented as points - are clustered around the mean of each dimension, and the dimensions intersect at their respective means. Thus, faces that are typical (i.e., not distinctive), are clustered near the middle of the multidimensional space, with distinctive faces existing further from the centre of the space and also surrounded by fewer points (i.e., other distinctive faces). For this reason, distinctive faces are easier to recognise, because they are surrounded by few competing points, but take longer to classify as face than typical faces, because they exist further away in Euclidean distance from the prototype of what constitutes a face.

Models that use PCA to explain face recognition extract the most relevant features for face recognition to simulate how faces are learned and recognised (e.g., Hancock, Burton, & Bruce, 1996; Turk & Pentland, 1991). To do this, PCA models reduce faces to eigenvalues (or eigenfaces), with the lowest eigenfaces representing the most variance, but higher eigenvalues represent more details. For efficiency, faces are then represented as a single value, which is the

[77] Other face recognition models exist, but these models do not attempt to explain how humans recognise faces. Instead, in these models, machines learn to recognise human faces without considering the theoretical architecture of human memory (e.g., Phillips, Moon, Rizvi, & Rauss, 2000). Thus, models that test machine recognition of human faces are not considered in the current thesis.

[78] The term, ‘vector space’, was introduced in Chapter 3 (see page 110) when discussing the simulated faces used for the two face-recognition experiments.

sum of the eigenvalues, from which they can be reconstructed. Using PCA, Hancock et al. (1996) were able to demonstrate that certain principal components underlie the psychological constructs, discriminability, memorability, and familiarity, and demonstrated that certain principal components were correlated with hits or false alarms.

Connectionist models map face recognition onto a network of neurons, which activate in response to stimuli. Of the four types of models, connectionist models and functional models are the only models that also represent associative memory, for example, faces and names (Burton & Bruce, 1992; Burton & Bruce, 1993), or faces and semantic information (Burton et al., 1990).

For a more detailed overview of the four types of models, see Rakover and Cahlon (2001).

Bruce and Young Model of Face Recognition

Bruce and Young (1986) proposed a theoretical model of face recognition aimed at explaining various aspects of person recognition. Their model is a functional model that does not include any computational simulations. Two theoretical parts constitute their model: The first part provides a broad explanation of how target faces are encoded, while the second part explains recognition of the target's face and subsequent access to the target's semantic information and name.

Pictorial and structural codes. In their model, Bruce and Young argue that there are two different types of information available in a face (i.e., pictorial and structural codes) that play a pivotal role in encoding. Pictorial codes are informational components that are picture-dependent and are not limited to the domain of faces, for example, a photograph of the Eiffel Tower would contain pictorial codes such as the lighting or weather. In contrast, structural codes are abstract components derived from the positional and dynamic relation of the facial features with each other, for example, the holistic information of the face, how the face moves during speech and when making expressions, and other information that is not view-dependent.[79]

Recognition of target faces. When observers are exposed to a face, they perceive and encode the individual components of the face (e.g., the mouth, or eyes). The degree to which individual facial components contribute towards recognition of familiar and unfamiliar faces, respectively, differs. Recognition of a familiar face, such as of a parent or partner, is still possible even if a component of the face is covered (Young, Hay, McWeeny, Flude, & Ellis, 1985; but also see Sadr, Jarudi, & Sinha, 2003), or if the component is presented to the viewer alone without the rest of the face (see Ellis, Shepherd, & Davies, 1979; but also see Kaufmann & Schweinberger, 2004). Everyday experiences suggest that recognition of familiar faces persists even as the face ages, facial adiposity changes, perspective changes, or some features change (e.g., hair style and hair colour). Contrastingly, changes to appearance or limited visual information are unlikely to result in recognition of unfamiliar faces.

Impaired recognition of unfamiliar faces, compared to familiar faces, may result from the limited encoding circumstance. Unfamiliar faces are typically encoded during fewer instances, which consequently provide the observer with limited opportunities to encode more structural information,[80] whereas familiar faces are encoded frequently and across different contexts, views, and time. The repeated exposure to familiar faces results in a richer, more diverse set of structural codes that facilitates recognition across various recognition conditions.

[79] In Chapter 3, I criticized the method used in research that tested recognition of pictures for not introducing any variation in the stimuli between encoding and recognition. My criticism derives from the distinction between pictorial and structural codes of information.

[80] The use of the terms ‘familiar’ and ‘unfamiliar’ are confusing within the face recognition literature. Typically, ‘unfamiliar’ faces belong to strangers, whereas ‘familiar’ faces belong to people whom we know ‘well’. An unfamiliar face, however, cannot be recognised if that face was never encoded – thus, is the face still ‘unfamiliar’, or is the face better described as ‘less familiar’? The use of these theoretical terms requires further refinement, but for the present purposes I will use the terms as they are used in the extant literature.

The model proposed by Bruce and Young is broken down into a set of hierarchical relations (see Figure 6.1). Whenever the observer sees a face, the observer encodes both the pictorial codes and structural codes. Since the extraction of structural and pictorial codes from a face occurs whenever the observer sees a face, Bruce and Young argue that face recognition opportunities behave as further encoding opportunities.[81] Structural information of the face leads to activation of a face recognition unit (FRU) for that face in memory, and the observer will experience a sense of familiarity towards the face. The connection between the FRU and the structural codes are strengthened through repeated exposure, so that faces are more easily (i.e., faster) recognised as familiar.[82]

[81] Nickerson (1965,1968) made a similar argument. Nickerson (1968) tested for single and doubled-encoded faces, which were designated 'New' and 'Old' faces, respectively, in Nickerson (1965). This is discussed in more detail in Chapter 3.

[82] Although it is not clear from the literature at what point an unfamiliar face becomes familiar.

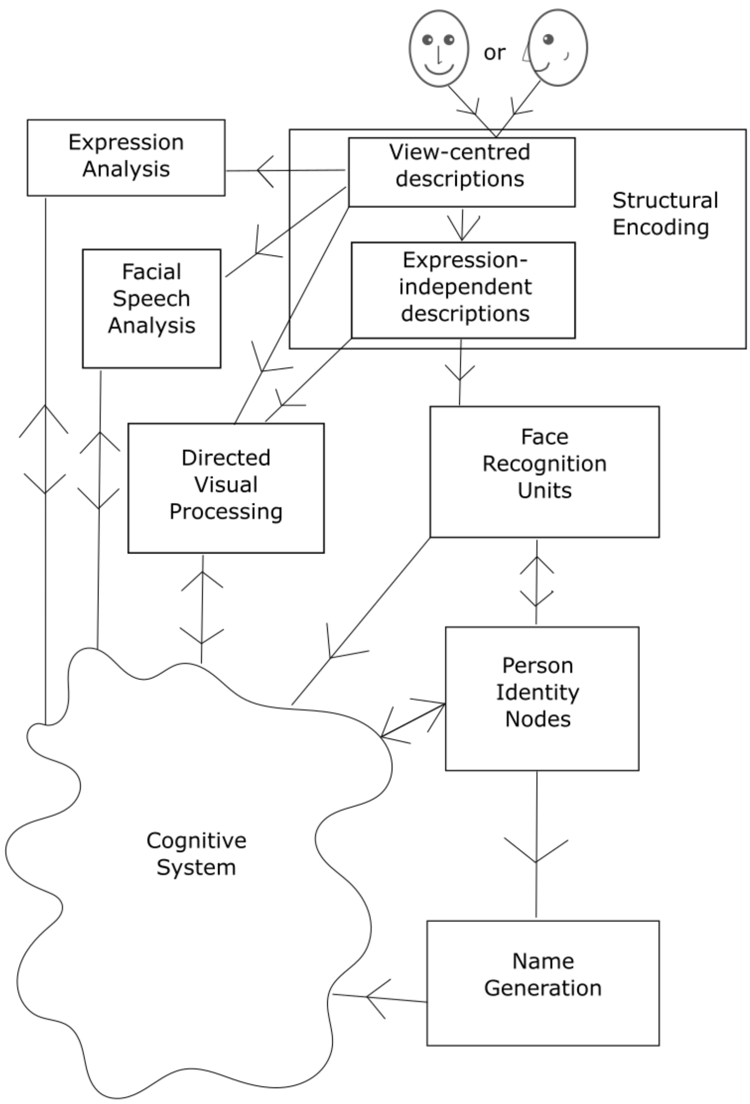

Figure 6.1. A reconstruction of the theoretical model of face recognition proposed by Bruce and Young (1986).

The sense of familiarity that arises from the activation of the FRU does not imply automatic retrieval of semantic and name information about the target. Instead, semantic information specific to that face is accessed via a second unit, the person identity node (PIN). Access to semantic information through the PIN occurs independently of, but prior to, access to name generation. Names are stored in a third unit, which activates last, and like the relationship between FRUs and PINs, name generation can only be accessed once the PIN has been activated. Due to the hierarchical structure of the Bruce and Young model, it is not possible to recall the target's name without first simultaneously accessing their FRU and PIN.

It is, however, possible to recognise a face as familiar and struggle to recall their occupation, or name. There is some support for the differential access of names and semantic information, as well as other face recognition errors. Young, Hay, and Ellis (1985) asked participants to keep a diary of the different memory errors that they experienced in daily life. The most frequently reported errors were (a) the individual was unrecognised, (b) the individual was misidentified, (c) the individual was familiar but no semantic information was recalled, and (d) the participant struggled to recall the semantic details of the individual. The results from Young et al. support the notion of a hierarchical relation between face recognition and recollection of names and semantic information.

Overall, the Bruce and Young model provides a solid theoretical framework to understand face recognition, and subsequent access to semantic and name information. Their model is one of the most widely-accepted and researched models of face recognition. There are, however, a few criticisms of the model (some suggested by Bruce and Young themselves). In some ways, the proposed model is quite crude. For example, Bruce and Young do not explain how face recognition units are populated for new, unfamiliar faces. Bruce and Young also do not specify whether the PIN is a gateway that allows access to semantic information, or if the PIN is where semantic information is housed. In a subsequent paper, Burton et al. (1990) designate the PIN as a pool that stores semantic information. Furthermore, the cognitive system, as shown in Figure 6.1, is undefined, and it is not clear what influence other cognitive processes (e.g., confidence) have on face recognition or how these processes interact with the PIN, FRU, and naming units in the current model.

The Bruce and Young model, however, was not intended to explain all aspects of face recognition. Instead, it was intended to be a starting point for broadly explaining how face recognition works, and to account for some of the observed memory errors reported within the research literature. One strength of the Bruce and Young model, which is pivotal to the current thesis, is that it does attempt to explain person recognition, that is the relation between the face of the target and the associated semantic information about the target.

Interactive Activation and Competition (IAC) Model of Face Recognition

In a later paper, Burton et al. (1990) used an artificial neural network to explain the causal mechanism behind associative priming. Associative priming is the phenomenon where a target face (e.g., Princess Diana) is recognised quicker when preceded by the presentation of a semantically associated face (e.g., Prince Charles) rather than a dissociated face (e.g., Walt Disney).

Interactive Activation and Competition models. To demonstrate associative priming, Burton et al. (1990) used a revised form of the Interactive Activation and Competition (IAC) network (McClelland, Rumelhart, & the PDP Research Group, 1986). In an IAC network, pools of units (i.e., neurons) exist, and each pool represents a type of information (see Figure 6.2). McClelland et al. (1986) explain the IAC architecture using characters from West Side Story: For example, one pool houses the names of the characters, another pool houses characters' ages, and another pool stores characters' professions (see Figure 6.2). There is also an 'instance' pool, which represents the 'identity' of each character by connecting each character's associated information. Units within pools are connected to units in other pools with bidirectional excitatory connections. When one unit is activated, it sends an impulse along the excitatory connection to units in other pools to activate them. Thus, the character, 'Art' is connected to the following other-pool units: pushers, 40's, Jet gang. When the unit 'Art' is activated, an impulse is sent to the units, '40's', 'pusher', and 'jet', causing them to activate in response. The bidirectional excitatory connections represent the 'activation' part of the model name.

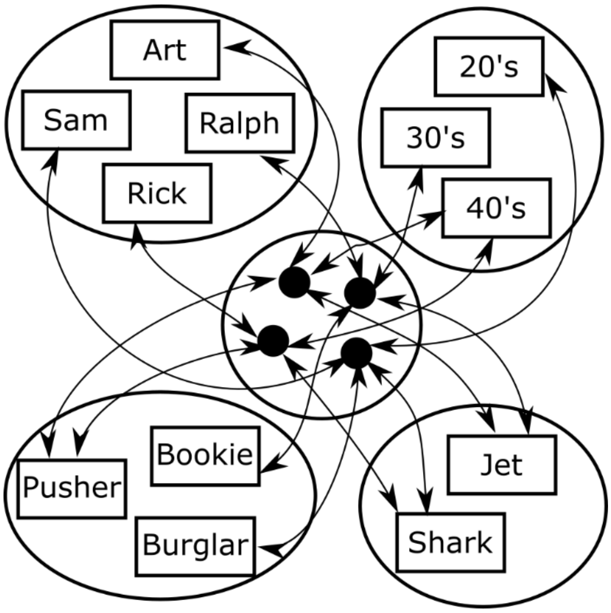

Figure 6.2. Revised version of the IAC model adapted from the PDP Handbook (McClelland et al., 1986). There are five pools in this diagram: A pool of names, ages, occupations, and gangs. The fifth pool is an 'instance' pool. Units are connected to units in other pools via the black arrows. Only excitatory connections are shown in this diagram; within-pool inhibitory connections are not shown.

Units within the same pool are connected to each other with inhibitory connections. When one unit within a pool is activated, it subsequently fires along its inhibitory connections to other intra-pool units and suppresses these units. When suppressed, these units are less able to activate. For this reason, when 'Art' is activated, the outgoing inhibitory connections from 'Art' suppress activation of 'Rick', 'Ralph', and 'Sam'. This constitutes the 'competition' part of the model name.

Computational explanation of the IAC update functions. The following section is added for readers who are interested in the technical details of IAC models. Calculating the activation levels of the units within the IAC model occurs over a variety of steps: net input, decay, and inhibitory input. I will briefly describe each step, and how the activation levels of the IAC models are calculated using a simple example of an IAC network (Figure 6.3), and referring to the explanations provided in McClelland (2015). The calculations that I am describing are not mathematical formulae, but are, instead, the computational algorithms used in the update function (in Matlab for McClelland et al., [1986], and in R for the current thesis).

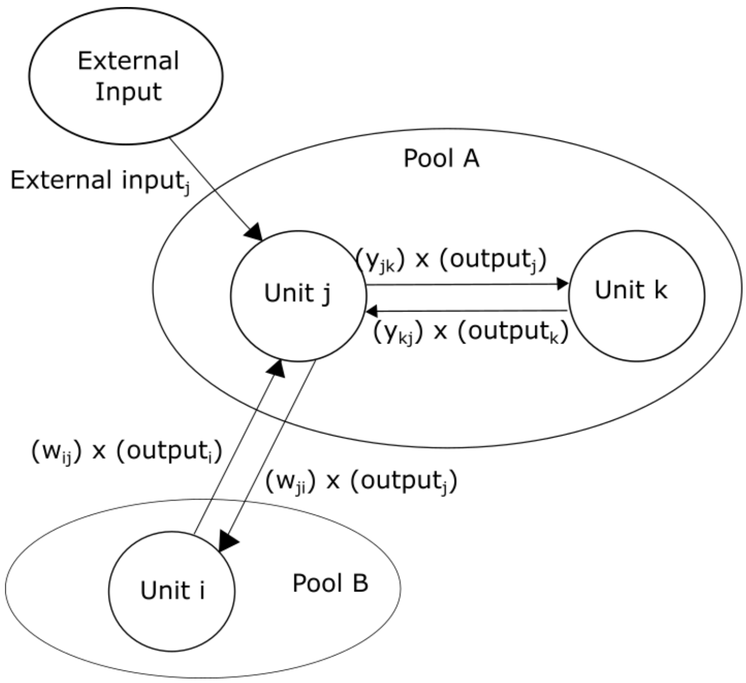

Figure 6.3. A demonstration of how units activate within an IAC network. Unit j receives excitatory input from an external source and from Unit i, but receives inhibitory input from Unit k. Unit k receives inhibitory input from Unit j. Unit i receives excitatory input from Unit j.

Parameters. All IAC models have a set of predefined parameters that remain constant. These parameters describe the minimum and maximum levels of activation, the activation at rest (i.e., the starting activation level), and the decay rate. Minimum is normally set to 0, and maximum to 1; however, all the parameters can differ.

Activation of units. Unit activation represents the degree of stimulation experienced by a neuron. Once units have activated to a satisfactory level, they will fire and provide activation to other connected units. For IAC models, the output of the unit is not binary (i.e., it is not 'on' or 'off'); instead, the output of the unit is graded, thus weaker activations result in weaker output signals.

The activation level of a unit is the summation of its excitatory and inhibitory connections. To illustrate, I will use Unit j in Figure 6.3. In Figure 6.3., Unit j exists together with Unit k in Pool A. Unit j has a connection to Unit i, which exists in Pool B. The connections between Unit j and Unit k are inhibitory because they exist in the same pool, but the connections between Unit j and Unit i are excitatory since they exist in separate pools. The strength of the connection between units is referred to as the weight. Unit j also receives input from an external source.

The weight of the connection from Unit i to Unit j is denoted as w_{ij}. The weight from Unit j to Unit i (denoted as w_{ji}) is not necessarily the same strength as w_{ij}. The inhibitory connections between Unit j and Unit k are labelled y_{jk} and y_{kj} respectively.

The weights of the entire network update through a complex and simultaneous interaction between activation and inhibitory connections described below.

Net input. The net input to any unit is the summation of all external sources of input, and all excitatory connections from units in other pools. The strength of the impulse sent from one unit to the next is the weight of the sending unit multiplied by the output of the sending unit.

The net input to Unit j is the summation of (a) the weight of its connection from Unit i, w_{ij} , which is multiplied by the output from Unit i, and (b) any other external input to Unit j. Therefore, there will be no effect if the output of Unit i is negative or zero, which implies that Unit i is not activated and did not fire. The formula for the net input to Unit j is:

$$net\ input_j = \left[\left(w_{ij}\ x\ output_j\right) + external\ input_j\right]$$

Inhibitory connections and competition. The net input is then reduced by the summation of all inhibitory connections. The strength of inhibitory connections is calculated by multiplying the weight of each sending unit with the output of that sending unit. For Unit j, the net input is reduced by the product of the strength of the connection from Unit k to Unit j, y_{kj}, and the output from Unit k. The net input equation is:

$$net\ input_j = net\ input_j -\ (y_{kj}\ x\ output_k)$$

In Figure 6.3, there is only one excitatory connection between Unit j and Unit i, but in theory, Unit j can receive input from multiple units simultaneously. In the same manner, Unit j can receive inhibitory input from multiple units simultaneously. Thus, the net input for Unit j is the summation of (a) the external input and (b) all products of the weights of all incoming units and their outputs. The net input to Unit j is then modified to account for the summation of all products of the weights of all inhibitory connections and their outputs. Simultaneously, the net inputs for Unit k and Unit i are also calculated and updated.

Decay. A unit cannot remain activated indefinitely, and so, without any input, the activation level will taper and slowly decay. The decay function will cause the activation of the unit to return to the predefined rest level. The change in the activation level of Unit j from the current cycle to the next, takes the effect of decay into the account. Two algorithms are proposed, which depend on the result of the next input. To calculate the change in activation levels of Unit j, the following formulae are used conditional on the value of the net input (McClelland et al., 1986). If the net input to Unit j is positive, then the following formula is used:

$$\Delta a_j = (maximum - a_j)net_j - decay(a_j - rest)$$

If, however, the net input for Unit j is negative, then the following formula is used:

$$\Delta a_j = (a_j - minimum\ )net_j - decay(a_j - rest)$$

The result of either equation is the change in the activation level of the receiving unit, in this example, Unit j. The activation level of Unit j is modified (i.e., updates) by the change in activation (Δa_j), which can be negative or positive.

Cycles. The model is run for a predefined number of cycles. Activation levels for all units within the model are calculated per cycle. There is no concrete, comparable time unit for a cycle: It is an abstract representation of some form of time or exposure. Activation levels for all units are stored per cycle, and activation levels can increase or decrease from one cycle to the next.

IAC model to explain associative priming. In the model proposed by Burton et al. (1990), information is not distributed in the same manner as in the Jets-Sharks model

(McClelland et al., 1986). Instead, the Burton et al. model is a blend of the IAC network (McClelland et al.) and logogen models (Morton, 1969). Logogen models are used to model word recognition, and within this type of model, each unit (or node) represents a 'collection' of information rather than a single item of information, like in the Jets-Sharks IAC model. In logogen models, words are recognised when the unit reaches a predefined activation level.

The face recognition model created by Burton et al. (1990) was moulded on the theoretical components of the Bruce and Young model of face recognition and consisted of three pools: a face recognition pool, a person identity pool, and a semantic information pool. Each pool consisted of its own respective units. Each unit within each pool is connected via bidirectional excitatory links to relevant, associated information within other pools; however, units within pool are connected to each other via inhibitory connections.

The aim of the model proposed by Burton and colleagues was to explain associative priming of faces. Their results showed that the activation of one face triggered partial activation of another face, if the faces were associated with each other. For example, the face recognition unit (FRU), person identity node (PIN), semantic information unit (SIU) of Prince Charles are connected to each other via excitatory connections; however, within the face recognition pool, the Prince Charles unit is connected to other face recognition units with inhibitory connections (see Figure 6.4). Each unit within the model represents a 'set' of information, for example, the FRU for Prince Charles represents the structural codes derived from his face, and the SIU represents the semantic information about him. The inclusion of such collections of information at the unit-level is the logogen component of the model.

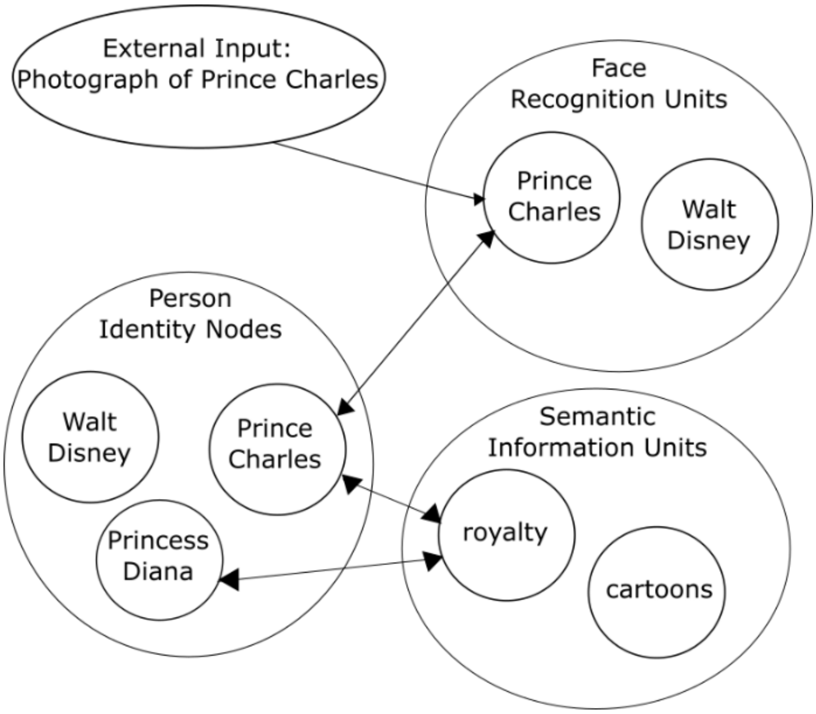

Figure 6.4. The revised IAC model constructed by Burton et al. (1990) to explain semantic priming. Inhibitory connections between same-pool units are not shown in this figure.

In their model, the 'association' between two faces is represented by excitatory links between each of the person identity nodes and the same unit in the semantic pool. Therefore, the person identity nodes for Princess Diana and Prince Charles have inhibitory connections with each other (because they exist within the same pool; these connections are not shown in Figure 6.4), but both nodes have excitatory connections with the same unit ('royalty') in the semantic information pool. Activation of the FRU and PIN for Prince Charles leads to activation of his semantic information (e.g. 'royalty'), which in turn leads to activation of the PIN for Princess Diana, although to a lower level. When the external stimulus (i.e., an image of Prince Charles' face) is removed, then the activation level of Prince Charles decreases, but only to the level of activation of the PIN for Princess Diana, which remains relatively stable. These two PINs (Prince Charles, Princess Diana) continue to activate each other, because they send activations to the unit 'royalty', which returns activations to both units. The simultaneous activation of two units results in a strengthened connection between them; this occurs through a Hebbian update function (i.e., a 'neurons that fire together, wire together' approach). Following this, when an external stimulus of Princess Diana is introduced to the model, her associated PIN activates quickly (because it was already partially activated) and reaches the arbitrary recognition level in fewer cycles (i.e., faster) than it took the Prince Charles PIN to reach the recognition level. Burton et al. (1990) designated the activation level of 0.45 as an indication of recognition.[83] It is through this process of pre-activation, due to the associative link of semantic information, that associative priming occurs.

One notable difference between the computational model of Burton et al. (1990) and the theoretical model proposed by Bruce and Young (1986) is the point within the models at which the cognitive process of 'recognition' occurs. In the theoretical model, Bruce and Young argued that various structural and non-structural information about the person (e.g., voice or gait) was stored in different pools, and that each pool led to recognition of that information type only. For this reason, they argued that face recognition occurred in the FRU pool when those units reached a predetermined threshold. In contrast, Burton et al. posited that recognition of a person results from a combination of various types of information, such as gait, speech, and face, and all this information feeds into a central unit for each 'person', which is the Person Identity Node. Burton et al. argued that recognition occurs at the PIN. However, activation of the PIN does not imply immediate access to the semantic information of that individual. It is for this reason that 'familiarity' of the person is perceived at the PIN level,[84] but access to the semantic information occurs further down in the hierarchy.

[83] This was an arbitrary level chosen by Burton et al. (1990). See page 367 of Burton et al. (1990).

[84] It is important to note the difference between face recognition and person recognition. Face recognition implies that the face is familiar, without the accompaniment of additional information about the person, whereas person recognition implies a 'deeper' recognition of the identity of that person, including recognition of the face, recognition/recollection of their semantic information, and possible access to their name. The disjunction between face and person recognition is like the distinction between 'remember' and 'know' judgements, where 'know' judgements are accompanied by a vague sense of familiarity, but 'remember' judgements are accompanied by a

IAC for Learning Unfamiliar Faces

One criticism of the IAC model proposed by Burton et al. (1990) is that it does not learn; in fact, the model is meant to demonstrate priming for familiar faces. In a subsequent paper, Burton (1994) adapted the IAC model so that it could learn new faces (known as the IACL model). The architecture of the IACL model is similar to the IAC model, but with some additions (see Figure 6.5). In the IACL model, a pool of Face Recognition Units exists, and each unit represents either a familiar or an unfamiliar face. In previous models, the external stimulus of a target face directly activates a FRU (on the condition that the target face was encoded before and the target face can be recognised from the external stimulus). In the IACL model, a preceding step occurs before FRUs are activated. In the IACL, several pools – each representing a type of feature – feed directly to the FRU pool. Each feature pool consists of various units of that feature (which vary perceptually).[85] All units within each feature pool are connected to each unit within the FRU pool, but the strengths of these connections differ. If an FRU represents a known face (e.g., the white FRU block in Figure 6.5), then it has one excitatory connection with a single unit in each feature pool. In contrast, if an FRU represents an unknown face (e.g., the grey FRU block in Figure 6.5), then it has connections with each unit in each feature pool, but these connections are assigned random, weak weights. The model learns which feature units are associated with an unknown face in the following manner: A unit is randomly selected from each feature pool, and this constellation of units fires, subsequently activates an FRU. After 35 cycles, the weights between the selected feature units and the FRU are updated using a Hebbian update function[86], and the process repeats. The same units are

recollection of other information (either episodic or semantic) associated with the encoded item (Yonelinas, 1994; Yonelinas, 2002).

[85] The feature pools and their respective units are meant to represent broad conceptual features. For the sake of demonstration, Burton (1994) described them as the features used in face composite systems, but recognised that these features may not exist in that form.

[86] With a Hebbian update function, units that activate together, or assist with each other's activation, have stronger excitatory connections. Thus, the weights between the randomly selected units in the IACL (burton, 1994) are strengthened because they fired together.

activated again (from rest), which leads to an activation of the connected FRU, and after 35 cycles, the weights of the connections are updated. After five trials, the activation level of the unknown FRU reaches the same activation level as an FRU representing a known face.

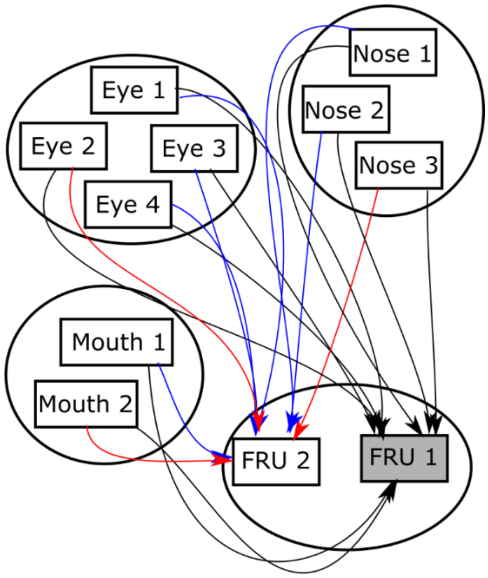

Figure 6.5. A revised and simplified version of the IACL model introduced by Burton (1994). This model represents how unfamiliar faces are learned. In the schematic, there are four pools. Three of the pools contain facial features, and the fourth pool is a face recognition unit that contains abstract representations of known and unknown faces. Each face is represented by a face recognition unit (FRU). The grey block in the FRU pool represents an unknown face, whereas the remaining two units in the same pool represent known faces. Red lines represent strong excitatory connections, and blue lines represent weak connections. Black connections are assigned random, weak weights.

The IACL model is innovative as it demonstrates how new faces are learnt through repeated exposure and strengthened memory trace; however, there are several shortcomings of the model. For example, all connections update in the same way, suggesting that features are of equal importance; however, there is evidence that some features are more salient than others (Ellis, Davies, & Shepherd, 1978; Goldstein & Mackenberg, 1966; Haig, 1984; McKelvie, 1976). In the same vein, perceptual characteristics such as distinctiveness and typicality are not represented in the IACL (a better model for this might be the face recognition models using multidimensional face space; Valentine, 1991). Burton (1994) also did not explain how the feature units came to exist within the pools. In the IACL the units are not treated as discrete units, but they exist in vector space instead. A final criticism of this model is that it demonstrates that new faces can be learned, but it does not demonstrate how new *people* – where faces and associated semantic information are connected – are learned.

Semantic versus Episodic Memory

For an eyewitness, the memory of the perpetrators' faces and all associated information (e.g., roles, actions) is embedded within the context of the episode (i.e., the crime). In the original model posited by Burton and colleagues, the PINs are connected to the units within a semantic information pool, and a sufficiently activated PIN allowed access to relevant semantic information about the familiar face. It is unlikely that the same structure used in the Burton et al. model (three pools: FRU, PIN, semantic information units) would suffice for eyewitness recognition for multiple perpetrators, as this would imply that roles are considered semantic information associated with the perpetrators. Is it possible that an eyewitness can recall the roles of the perpetrators (semantic recollection) *without* any recollection of the event (episodic recollection)? This requires a more formal investigation of the relationship between semantic and episodic memory.

Tulving (1972) first proposed that two memory systems exist (episodic, and semantic[87]), and later added three more memory systems (procedural, perceptual representation, and short-term memory; these three systems are not of interest to the current thesis). Broadly, episodic memory refers to experienced events, and semantic memory refers to abstract facts, but Tulving further described semantic memory as referring "to the speaker's 'knowledge' rather than his 'remembering'" (Tulving, 1972, p.387). The two types of memory are accompanied by different conscious experiences (Tulving, 1985a; Tulving, 1985b). Episodic memory is accompanied by autonoetic consciousness, which is analogous to the conscious experience accompanying 'Remember' responses. Contrastingly, semantic memory

[87] Although the term 'semantic memory' was first used by Quillian (1966, as cited in Tulving, 1972). Tulving proposed that 'episodic memory' should exist as the antithesis to semantic memory, much like verbal and nonverbal memory are contrasted with each other.

is accompanied by noetic consciousness that are akin to the conscious experience accompanying ‘Know’ responses.

Tulving (1985a, 1985b)revised the relationship between episodic and semantic memory, arguing that they do not exist independently of each other. Instead, Tulving proposed that episodic memory developed from semantic memory, and thus is an extension of semantic memory. Furthermore, Tulving argues that the dependence between semantic and episodic memory is unequal: Semantic memory can be stored and retrieved without episodic memory, but episodic memory cannot exist without semantic memory. Semantic and episodic memory are often treated as though they are completely distinct from one another, but they are not: All memory originates from episodes, and thus, at some stage, all memories have an episodic component. Baddeley suggests a similar relationship between the two memory types when he states that they may rely “on a common episodic input system” (Baddeley, 2001, p. 1346). In fact, this relationship between episodic and semantic memory is suggested in the following definition provided by Bower (2000):

> “In these terms, the memories typically being acquired and tested in verbal learning experiments are episodic (e.g., “I just studied the pair king-table). Semantic information are the abstracted words, concepts, and rules stored in our long-term memory whose context of acquisition was long ago forgotten (e.g., most people cannot say when they first learned the meaning of the word king).” (Bower, 2000, p. 22)

Bower’s definition suggests a hierarchical relationship between semantic and episodic memory, but it is not clear whether semantic memory is a weaker or stronger form of memory compared to episodic memory. There are a few possible ways that semantic memory can emerge from episodic memory. First, the “context of acquisition” is forgotten because the memory of the episodic information decays, whereas the semantic information is retained. Second, the semantic information is reinforced repeatedly, resulting in strengthened semantic connections but weak, *unchanged* episodic connections. With time, the strengthened semantic connections overtake the episodic connections, thus making it easier to recall the semantic information without remembering where the information was learnt. Third, the same semantic information is reinforced across *multiple, different* episodes, resulting in an equally weak set of links between various episodes that also inhibit one another, but a stronger, reinforced set of connections with the semantic information.[88] It is not clear which, if any, of these options is valid, but it is certain that memory must begin with an episodic connection. It is not the aim of this thesis to determine whether semantic or episodic memory exist, or whether the taxonomy is correct. Instead, it is important to note that the IAC model proposed by Burton et al. (1990) included the relationships between familiar faces and semantic information, and it is not clear whether the same model can be used for eyewitness recognition of unfamiliar faces and episodic memory about the crime. Furthermore, it is unclear whether role pairing would be considered semantic memory about the perpetrator or episodic memory about the crime.

Aim and Rationale

The aim of the current model was to provide a computational explanation for eyewitness difficulties with recognising multiple perpetrators. To do this, the IAC model used by Burton et al. (1990) was adapted for the current experiment. However, unlike the original IAC model which tested recognition of familiar faces and their related *semantic* information (Burton et al., 1990), the revised model tested facial recognition of multiple perpetrators and their related roles – a test of *episodic* material.

Method

Structure of the revised IAC models

For the current study, three different architectures were proposed for the IAC model with the aim of testing the failure to link roles with faces (see Figure 6.6). All three architectures have the following similarities:

[88] Baddeley seems to suggest this as well, when he states, "it seemed simpler to assume that semantic memory merely represents the residue of many episodes" (Baddely, 2001, p. 1346).

- The models consist of four pools: Face Recognition Units, Person Identity Nodes, Crimes, and Roles, and a hidden layer, which cannot receive external input. The hidden layer is not relevant to this discussion, and is not shown in Figure 6.6.[89]
- Each pool consists of ten units. All within-pool units are connected to one another with inhibitory links that are not displayed in Figure 6.6.
- The Face Recognition Units (FRU) represent the cognitive model of the physical description or mental image of each perpetrator. Each Face Recognition Unit is linked to a specific unit within the pool of Person Identity Nodes. When a FRU activates, it will also activate the respective node within the Person Identity Nodes pool.
- Activation of the PIN indicates 'recognition' (but not necessarily 'recall') of the respective individual. In all three models, the PIN pool allows access to the Crime pool.
- The Crime pool represents the episodic information about the crime. The Role pool represents the unique, episodic information about the role that each perpetrator performed within the crime. The connections between the PIN, Crimes pool, and Roles pool were manipulated within the three different models, and will be discussed in more detail later.
- Within all three models, each face recognition unit has a unique, excitatory link with a specific person identity node unit. For example, FRU-1 (which is the face of Perpetrator 1) is linked to Person Identity Node-1 (also known as PIN-1), and FRU-2 is linked to PIN-2, and so forth.

[89] The hidden layer receives input from the other pools, but it does not send any output to any other pools. Thus, the activation of the hidden layer units remains unchanged across all 100 cycles. I retained the hidden layer from the compANN package, but it plays no part in the model.

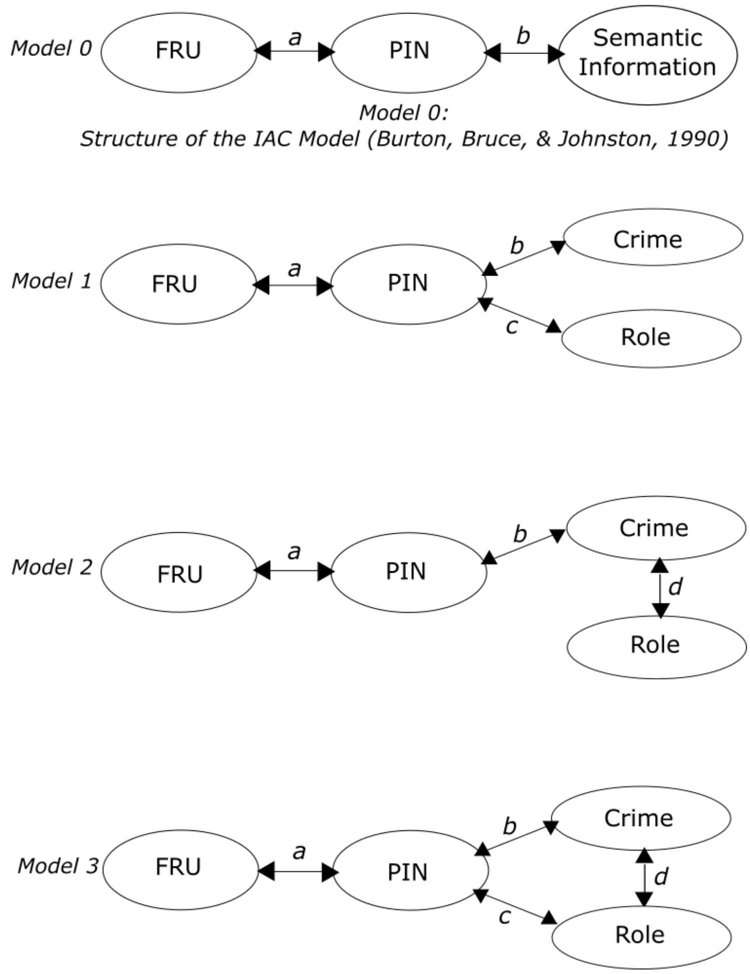

Figure 6.6. A visual representation of the three models constructed in the current experiment. FRU denotes Face Recognition Unit pool, and PIN denotes Personal Information Node. Model 0 is the architecture of the IAC model proposed by Burton et al. (1990). All three subsequent models retain the FRU pool and the PIN pool. Unlike the Burton et al. (1990) model that includes a Semantic Information Pool, the subsequent models include a context (‘Crime’), and the roles performed within the context (‘Role’). The difference between Models 1, 2, and 3 is the number of paths included in the model. In Model 1, the PIN units are connected to the Crime and Role pools separately, but the Crime and Role pools are not connected to each other directly. In Model 2, the PIN pool is only connected to the Crime pool, which connects to the Role pool. For Model 3, the PIN, Crime pool, and Role pool are all connected to one another.

In the IAC model proposed by Burton and colleagues (Model 0 in Figure 6.6), there are associative links between the face recognition units and semantic information. For example, Burton et al. (1990) could ask the participant whether ‘Al Capone’ was a ‘gangster’. But, as mentioned, the nature of the associations between different types of information may change for eyewitnesses. If the model was reworked so that it mimicked that of Burton et al., then both crime and role would be treated as two separate semantic pools, each accessed separately by the PIN and not connected to each other (Model 1, as shown in Figure 6.6).

However, role could be considered an attribute of the crime (Model 2, as shown in Figure 6.6), because an eyewitness might not be able to recall the role performed by the perpetrator without recalling the general nature of the crime as well. Two criticisms can be levelled at Model 2: First, the crime pool will become activated regardless of which perpetrator is seen, and this activation will, in turn, lead to an activation of *all* roles. The implication is that eyewitnesses will never be able to recall which perpetrator did what, and this performance will be *equally* poor across *all* set sizes. This seems unlikely, since the results from the previous studies reported in this project have shown that performance decreases as a function of Set Size. The second criticism is that the construction of the episodic event should be a prototype for how all episodic events are constructed, and should not be limited to only crimes. If we presume that Model 2 is the appropriate prototype, then this does not allow for episodic information to be lost while semantic information is retained, as suggested by Bower (2000).[90]

To address both criticisms, Model 3 is proposed where Role is considered an attribute of a crime, but is also an attribute of the perpetrator. Therefore, when a PIN is activated, both its respective Crime and respective Role will be activated. The activated Crime will provide additional activation to all roles that were performed within the crime.

A single-perpetrator version of all three models was built. In the single-perpetrator version, each crime was committed by one perpetrator, and there was only one role.

Scripts and Packages in R

All the models were constructed in R Studio (2014) using an adapted version of compANN package (Meyer, 2014), which measured machine responses to stimuli that varied

[90] One can argue that an eyewitness, or witness, will never forget the episodic experience of a crime, and for this reason, this memory does not need to conform to this suggestion. It is possible that these types of memory events are extremely distinctive, flashbulb-type memories that are encoded in such a way that they will never be forgotten. The consideration of the longevity of memory for different events extends beyond the scope of the thesis.

randomly in shape (e.g., triangle, square) and colour. The `compANN` package was written in C++, and included the Grossberg variation of the update function.[91] The basic update function used in the `compANN` package was translated into R, and the update function was edited to remove the Grossberg variation, because Burton et al. (1990) did not use the variation. Additionally, the plotting functions were updated. Two functions were written for the current experiment. The first function listed all the variations of the matrix weights, and the second function included the update functions. All functions can be found via the online link.

The parameters (minimum, maximum, rest, and decay) as well as the inhibitory and excitatory weights were set to the same values as those used by Burton et al. (1990). These values are listed in Appendix P.

Results

Testing the Three Models Against One Another

To determine which of the three models (1, 2, 3) best explained the results for the single-perpetrator scenario, I set the external input for each model to Perpetrator 1 and ran each model through 100 cycles. Like Burton et al. (1990), I used the same activation level of 0.45 to indicate recognition.

The results for each of the three models are listed in Table 6.1. Of the three models, Model 3 was the only model where Crime and Role reached the recognition threshold of 0.45 (Table 6.1). After 100 trials, Crime and Role reached activation levels of 0.32 and 0.32, respectively, for Model 1, and 0.39 and 0.23, respectively, for Model 2. The experiments reported in the current thesis show that eyewitnesses of single-perpetrator crimes do not forget the role of the perpetrator. For this reason, neither Model 1 nor Model 2 are supported by the results of the current thesis.

[91] Grossberg (1980) introduced a variation to the update function for the IAC model. His variation is a single update equation that includes the excitatory connections, inhibitory connections, and the decay function. This is unlike the decay formula used by McClelland et al. (1986) where two possible decay functions can be used and thus is preceded by an *if* statement.

Table 6.1

Activation Levels of Face Unit 1, PIN Unit 1, Crime Unit 1, and Role Unit 1 Across Three Models

		FRU Pool		PIN Pool		Crime Pool		Role Pool	
Model	Input	Unit	Activation	Unit	Activation	Unit	Activation	Unit	Activation
1	Face 1	FRU 1	0.68	PIN 1	0.55*	Crime 1	0.32	Role 1	0.32
2	Face 1	FRU 1	0.68	PIN 1	0.49*	Crime 1	0.39	Role 1	0.23
3	Face 1	FRU 1	0.68	PIN 1	0.62*	Crime 1	0.51*	Role 1	0.51*

Note. Asterisks (*) denote activation above 0.45, implying recognition. Each model was run for 100 cycles.

Manipulating Set Size for Model 3

Basic architecture of variations of Model 3. The basic architecture of Model 3 is shown in Figure 6.6. Model 3 comprises four pools: a face recognition units pool, person identity nodes pool, crimes pool, and units pool. Within each pool, there are ten units. All units within the same pool have inhibitory connections with each other.

Five variations of Model 3. Five versions of Model 3 were created, where each version represented a different crime scenario that was committed by one, two, three, five, or ten perpetrators. All five versions have the same architecture, and the only difference between the them is the excitatory connections.

In the single-perpetrator version of Model 3, there are excitatory connections between Face 1, Person 1, Crime 1, and Role 1. This is displayed in Figure 6.7. In the ten-perpetrator version of Model 3, each Face is linked to one Person via an excitatory connection (e.g., Face 1 is linked to Person 1, Face 2 is linked to Person 2, and so forth). There are also excitatory connections from each Person to a specific Role via an excitatory connection (e.g., Person 1 is linked to Role 1, Person 2 is linked to Role 2, and so forth). However, all Persons, and all Roles, are linked to only one crime (Crime 1). This is illustrated in Figure 6.8.

The retention of the other units within both diagrams are to illustrate the difference in how the excitatory connections are designed in the single-perpetrator and the ten-perpetrator versions of Model 3 (compare Figure 6.7. and Figure 6.8). Thus, the only difference between the five versions of Model 3 are the excitatory connections between units. Even though it would be strange for an eyewitness to have ten crimes in memory, Figure 6.7 and Figure 6.8 are meant to provide a visual demonstration of how the relations between the units are structured within single-perpetrator and ten-perpetrator versions of the model. [92]

Each model was named after the number of perpetrators involved in the crime, for example, Model 3.1 is the single-perpetrator version, and Model 3.10 is Model 3 where the crime was committed by 10 perpetrators.

Assumptions of the architecture. The architecture of Model 3 has numerous assumptions. All excitatory connections had the same weights (1.0), and all inhibitory connections had the same weights (-0.1). These are the same values used by Burton et al. (1990). The implication of using the same values is the assumption that all three pools (perpetrators, crimes, roles) are encoded with the same memory strength, and that units within pools are encoded with the same memory strength. It is possible to alter the weights to be any value, but the sheer number of possibilities increases the complexity of the model. Consequently, I decided to keep the values constant as I had no theoretical reason to change them.

A second assumption of the model is that the roles are completely distinct from the crime. The original reason for two distinctive pools for crimes and role was to account for the episodic origin of all memory, but also allow for the possibility that episodic memory deteriorates/strengthens into semantic memory. [93] Asking participants for a laboratory

[92] This may appear strange, but I have retained all the units in the diagrams and the R script for two reasons. Figures 6.7 and 6.8 demonstrate the relation between the relevant units for their scenarios, and act as a visual demonstration of the matrices that I constructed in R. I retained all the units within each of the five models, so that the size of matrix did not change across versions. The excitatory and inhibitory connections are set within a matrix (i.e., a 50-row x 50-column matrix). If the size of the matrix changed, then most of the R code would have to be updated to account for these varying dimensions, and increased the risk of human error. The retention of the units across the versions of Model 3 does not impact the output, because the units that are not relevant to that version do not receive input.

[93] This is dependent on how semantic memory develops from episodic memory.

experiment to recall the role of the perpetrator might be considered semantic memory; however, I question whether this is semantic memory. Instead, I wonder if this might be episodic memory.

A consequence of the architecture used for the five versions of Model 3 is the absurd nature of how single-perpetrator crimes would be encoded - specifically, whether roles and crimes are truly distinct. This was also a concern for the eyewitness experiment presented in Chapter 5. It seems unlikely that an eyewitness to a single-perpetrator crime would forget the role – in fact, the role is encapsulated within the nature of the crime. It for this reason that role recollection is not problematic for single-perpetrator crimes. However, removing the pools of roles from the single-perpetrator condition – while retaining this pool for the multiple-perpetrators conditions - is not wise. The architecture (i.e., structural design) of the multiple-perpetrator versions and the single-perpetrator version should not differ from each other, because this would confound the results of any simulations.

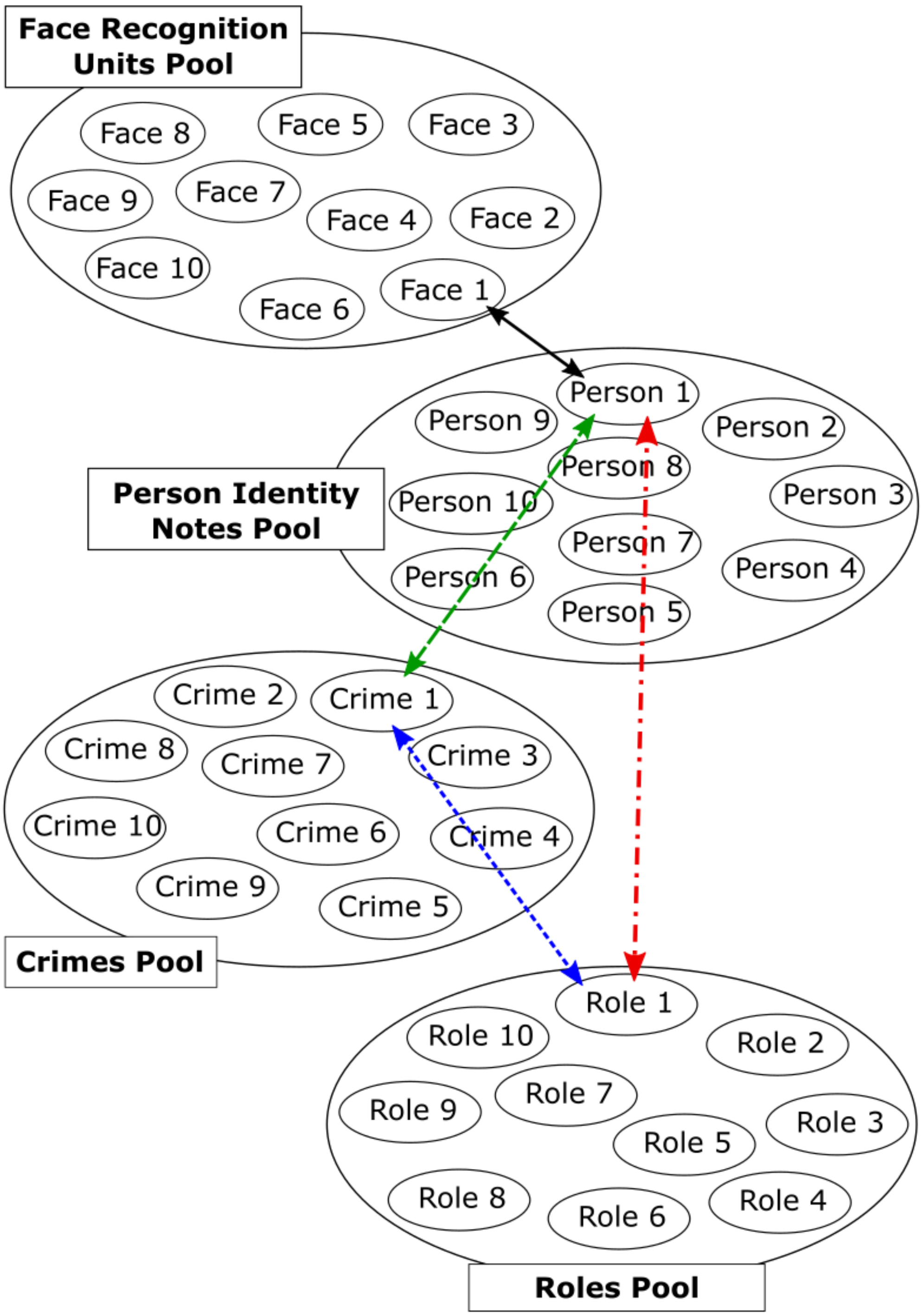

Figure 6.7 A representation of the four pools in the reworked IAC model for eyewitness memory for multiple perpetrators. This figure shows how the model is set up for a single-perpetrator crime. Face 1 is linked to Person 1, who is linked to Crime 1. Person 1 and Crime 1 are both linked to Role 1. Units within the same pool are connected to each other with mutual inhibitory connections that are not displayed here. Assume that Face 2 is connected to Person 2 who is linked to Crime 2, where Perpetrator 1 committed a unique role (Role 2). The colour of the arrows show the connections between PINs and crimes (green arrows), PINs and roles (red arrows), and crime and roles (blue arrows).

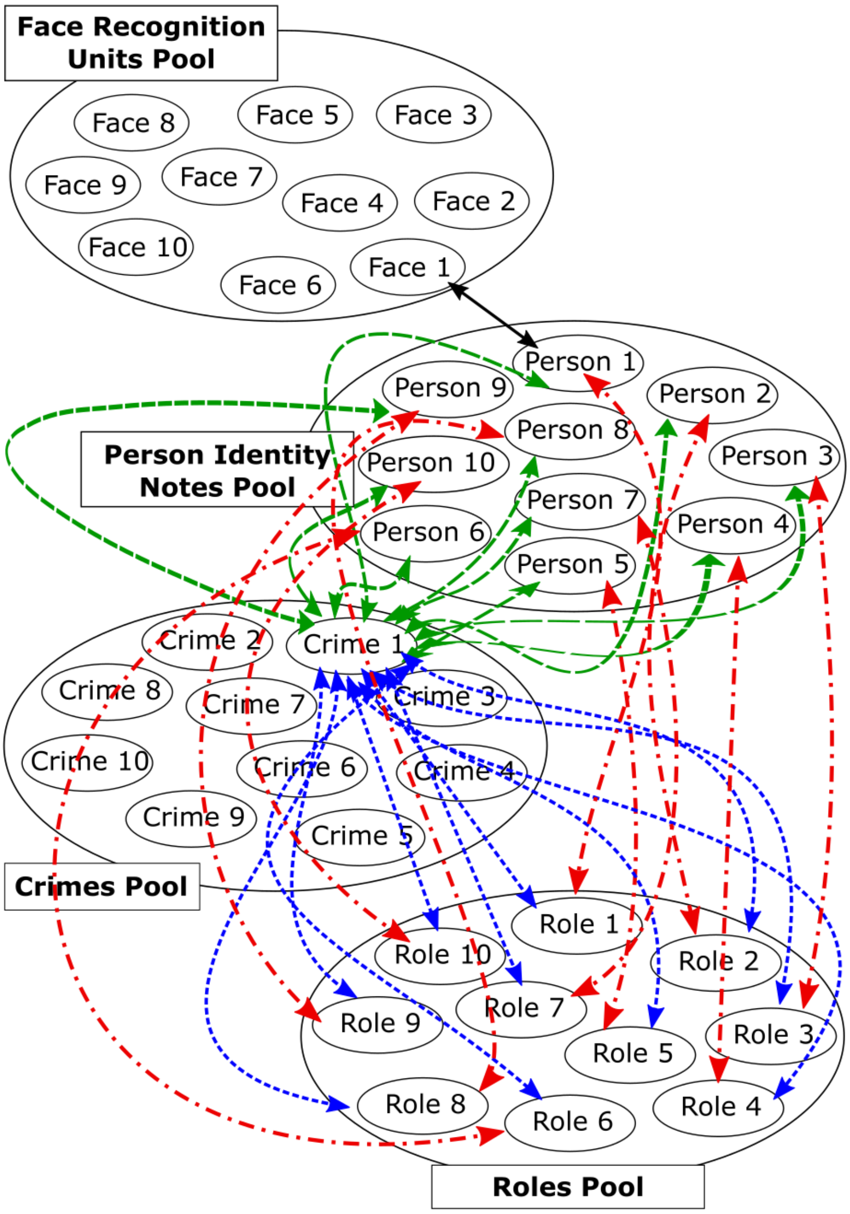

Figure 6.8. This figure is a representation of the reworked IAC for multiple perpetrators and demonstrates the ten-perpetrator condition. In this model, each face is linked to a unique perpetrator. All ten units in the Person Identity Nodes pool (i.e., perpetrators) are linked to a single crime, Crime 1, in the Crime pool; however, each person performed a single, and unique role. Therefore, each role in the Roles pool is connected to (a) their respective perpetrator, and (b) the crime in which they were performed. Units within the same pool are connected with mutual inhibitory connections, which are not shown here. For demonstration purposes, Crime 2 is not linked to any perpetrators nor roles; this has no effect on the outcome on the model. Assume that Face 2 is connected to Person 2, Face 3 is connected to Person 3, and so forth. The colour of the arrows show the connections between PINs and crimes (green arrows), PINs and roles (red arrows), and crime and roles (blue arrows).

Simulating Recognition for Perpetrator 1 and Associated Information

A simulation was run using each of the five versions of Model 3. In each simulation, the face of Perpetrator 1 was entered as external stimulus into each of the model versions. The external stimulus of Perpetrator 1 activated FRU 1 (i.e., the Face Recognition Unit linked to Perpetrator 1). The simulation ran for 100 cycles, and the activation levels for FRU 1, PIN 1, Crime 1, and Role 1 are shown in Table 6.2.

Table 6.2.

Activation Levels for Face Recognition Unit (FRU) 1, Person Identity Node (PIN) 1, Crime 1, and Role 1 Across Five Versions of Model 3

Model version	Number of perpetrators who committed the crime	Activation Levels After 100 Cycles			
		FRU 1	PIN 1	Crime 1	Role 1
3.1	1	0.68	0.62*	0.51*	0.51*
3.2	2	0.68	0.64*	0.69*	0.55*
3.3	3	0.68	0.64*	0.77*	0.55*
3.5	5	0.68	0.64*	0.85*	0.54*
3.10	10	0.68	0.61*	0.92*	0.50*

Note. Asterisks (*) denote activation above 0.45, implying recognition. These simulations were run on the five versions of Model 3. The input to each model was the face of Perpetrator 1, and each simulation was run for 100 cycles.

Based on the five simulations shown in Table 6.2, PIN activation of PIN 1 remained relatively stable as Set Size increased. The activation levels for Crime 1 increased as set size increased, but this was due to the increased number of connections between Crime 1, Roles, and Perpetrators. Crime 1 received more activation in the higher set sizes groups than the lower set size groups due to the increased number of excitatory connections. The activation levels for Role 1 remained relatively stable across the five simulations, and returned to almost the same level in Set Size 10 as in Set Size 1. In conclusion, what the simulation shows is that the poor recognition performance demonstrated by eyewitnesses for multiple-perpetrator crimes is not due to reduced activation of the relevant PIN and its respective Crime, and Role. Thus, these model simulations suggest that the poor recognition performance demonstrated by eyewitnesses to multiple-perpetrators is likely not due to a memory deficit, since all units reached an activation of 0.45.

Plotting Recognition for All Units

Across the five simulations, the activation levels for all PINs, Crimes, and Roles were plotted onto separate graphs (Figure 6.9, Figure 6.10, and Figure 6.11 respectively). Each simulation (1, 2, 3, 5, or 10 perpetrators) is labelled '3.1', '3.2', '3.3', 3.5', or '3.10' respectively in each figure.

Activation of Person Identity Nodes. Inspection of Figure 6.9 shows that PIN 1 (which is linked to Perpetrator 1) activates the most across the five simulations. This result is due to the direct connection between PIN 1 and FRU 1, which is immediately activated by the external input of the face of Perpetrator 1.

In scenarios where multiple perpetrators committed the same crime, PIN 1 was most activated unit of all the PINs. PIN 1, however, was accompanied by lesser activations of the PINs associated with the other perpetrators who are linked to the same crime. Therefore, in Panel 3.2 (the two-perpetrator crime), PIN 2 activated as well, albeit to a lesser extent than PIN 1. The pattern of lesser activation of associated PINs continued across all five simulations.

The activation levels of the associated PINs all reach the 'recognition' threshold of 0.45, but decrease slightly across set size. In the two-perpetrator crime (Panel 3.2 of Figure 6.9), PIN 2 had an activation level of 0.58, but this decreased to 0.51 in the ten-perpetrator condition (Panel 3.10 of Figure 6.9). The associated units had little impact on PIN 1, whose activation remained stable in Panels 3.2, 3.3, and 3.5. It was only in the ten-perpetrator condition that the associated units had an inhibitory effect on PIN 1 by decreasing its activation level from 0.66 (in Model 3.2) to 0.61 (in Model 3.10).

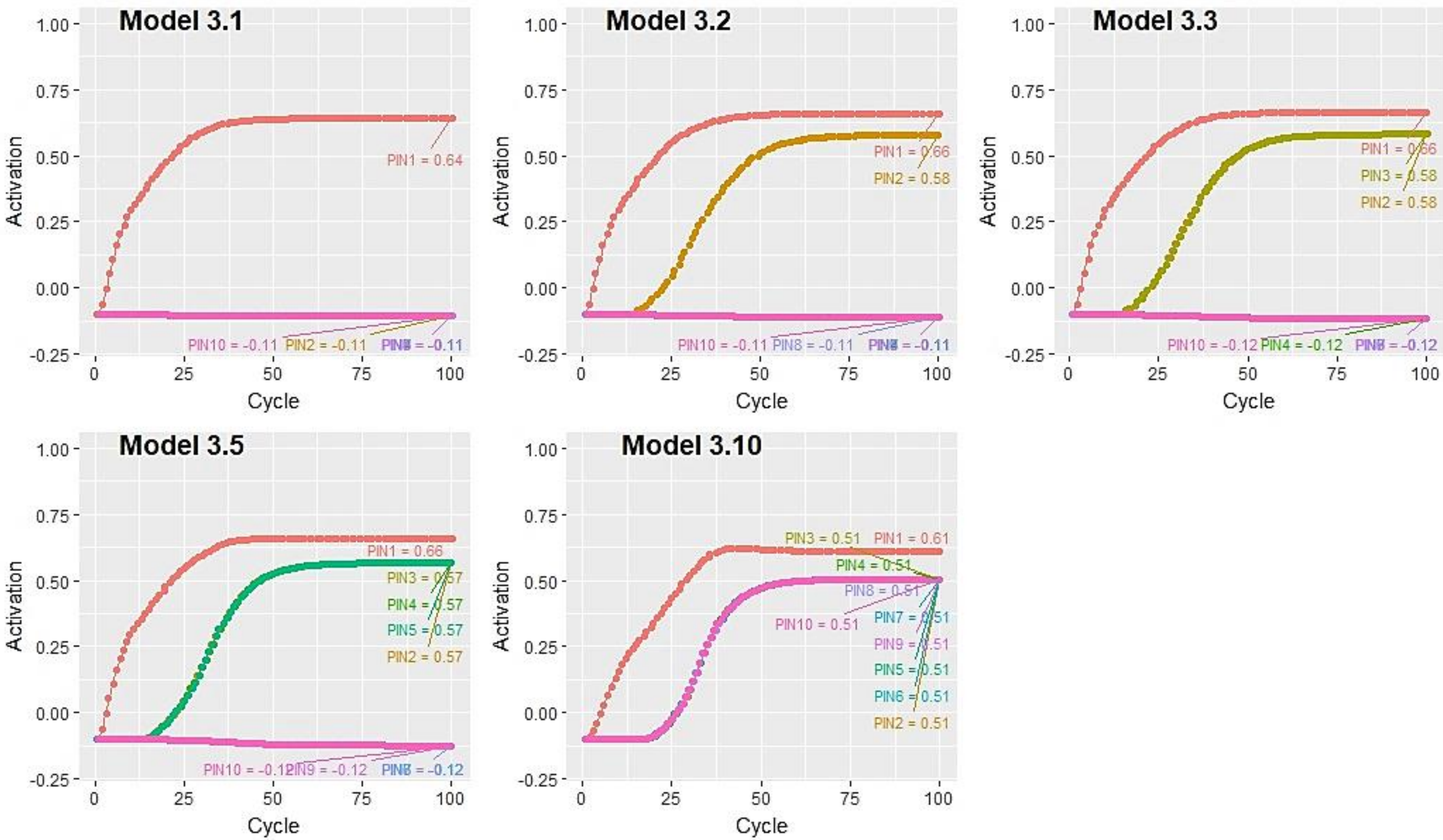

Figure 6.9 Activation levels for Person Identity Nodes (PIN) across the five models after 100 cycles. The name of each model is listed in each panel (3.1, 3.2, 3.3, 3.5, and 3.10). The input for each scenario was the face of Perpetrator 1. PIN 2 refers to the Person Identity Node of Perpetrator 2, and so on. The recognition threshold is set to an activation level of 0.45.

Activation of Crime Nodes. The activation levels of Crime 1 (Figure 6.10) replicated the data in Table 6.2. Crime 1 was always the most, and only, activated unit within the Crime pool, and this was due to the excitatory input from all the connected PINs. Crimes committed by an increasing number of perpetrators had a higher activation level than single-perpetrator crimes. The increasing activation level across Models was a consequence of the Crime unit in the multiple-perpetrator scenarios receiving increased excitatory input from the increased number of PINs and Roles. Thus, the activation levels of Crime 1 increased as set size increased: Crime 1 reached activation levels of 0.51 and 0.92 in Set Size 1 (Panel 3.1) and Set Size 10 (Panel 3.10), respectively.

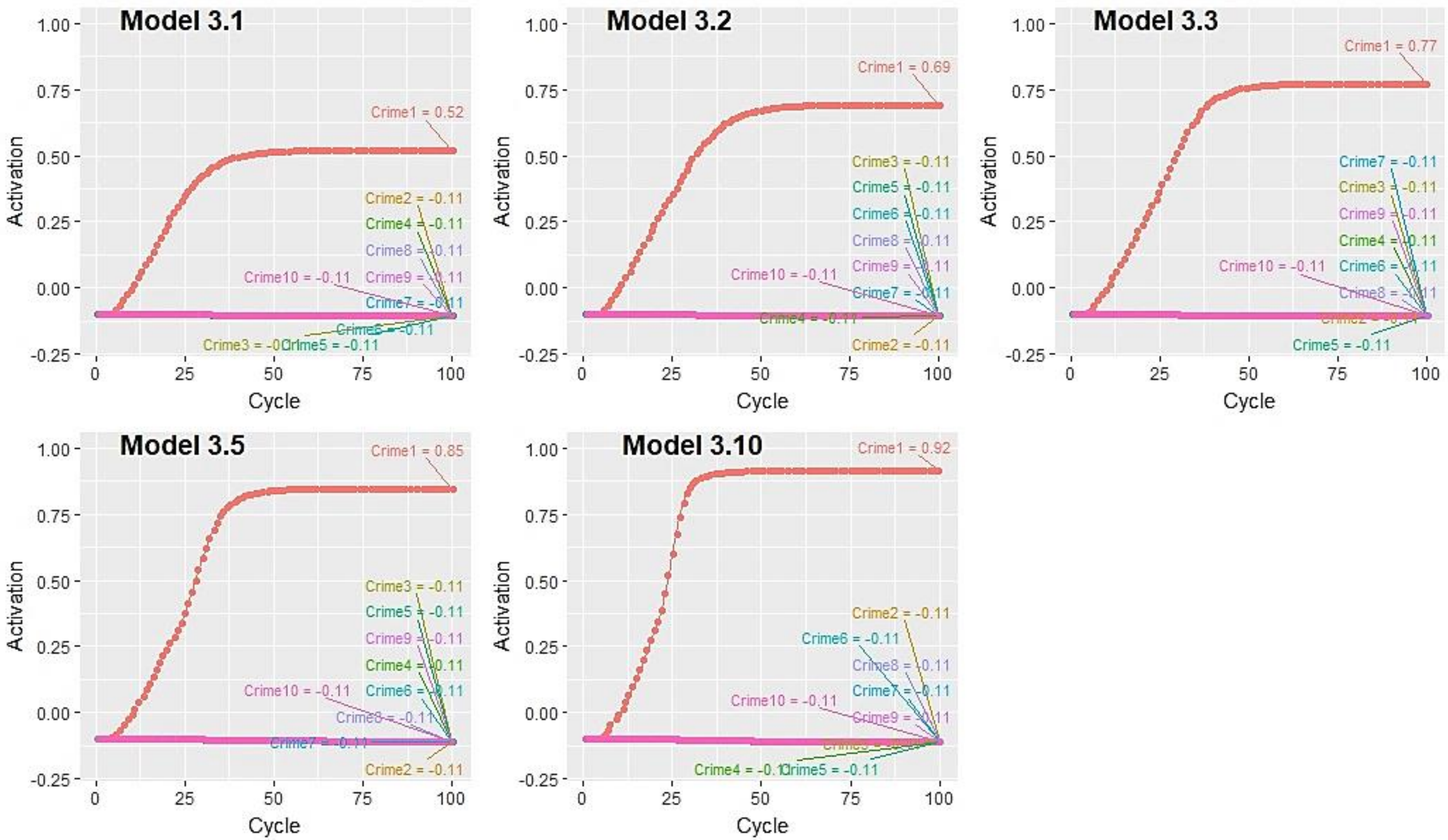

Figure 6.10 Activation levels for Crime nodes across the five models after 100 cycles. The name of each model is listed in each panel (3.1, 3.2, 3.3, 3.5, and 3.10). The input for each scenario was the Perpetrator 1. Crime 1 is the key crime that was committed by the designated number of perpetrators in each condition. Recognition is set to the activation level of 0.45.

Activation of Role Units. The activation levels for the Role Units across the five simulations are shown in Figure 6.11. Like the results for the PIN activations, Role 1 activated the most across the five simulations; however, the increase in Role 1 activation does not appear linear. Surprisingly, the activation of Role 1 increased slightly from the one-perpetrator to two-perpetrator condition (Panels 3.1 and 3.2 in Figure 6.11), and started to decrease in the five-perpetrator condition (Panel 3.5 in Figure 6.11).

In scenarios where more than one perpetrator committed a crime, Role 1 was accompanied by activations of associated roles (Panel 3.2 – 3.10), and these roles achieved the recognition threshold (all activation levels higher than 0.45). Compared to the two-perpetrator (0.53) and five-perpetrator scenarios (0.53), there was a minor increase in activation levels for the associated units in the three-perpetrator scenario (0.54). Compared to the other multiple-perpetrator conditions, the roles for the other associated perpetrators had the lowest activation in the ten-perpetrator condition (activation = 0.48; Panel 3.10 of Figure 6.11).

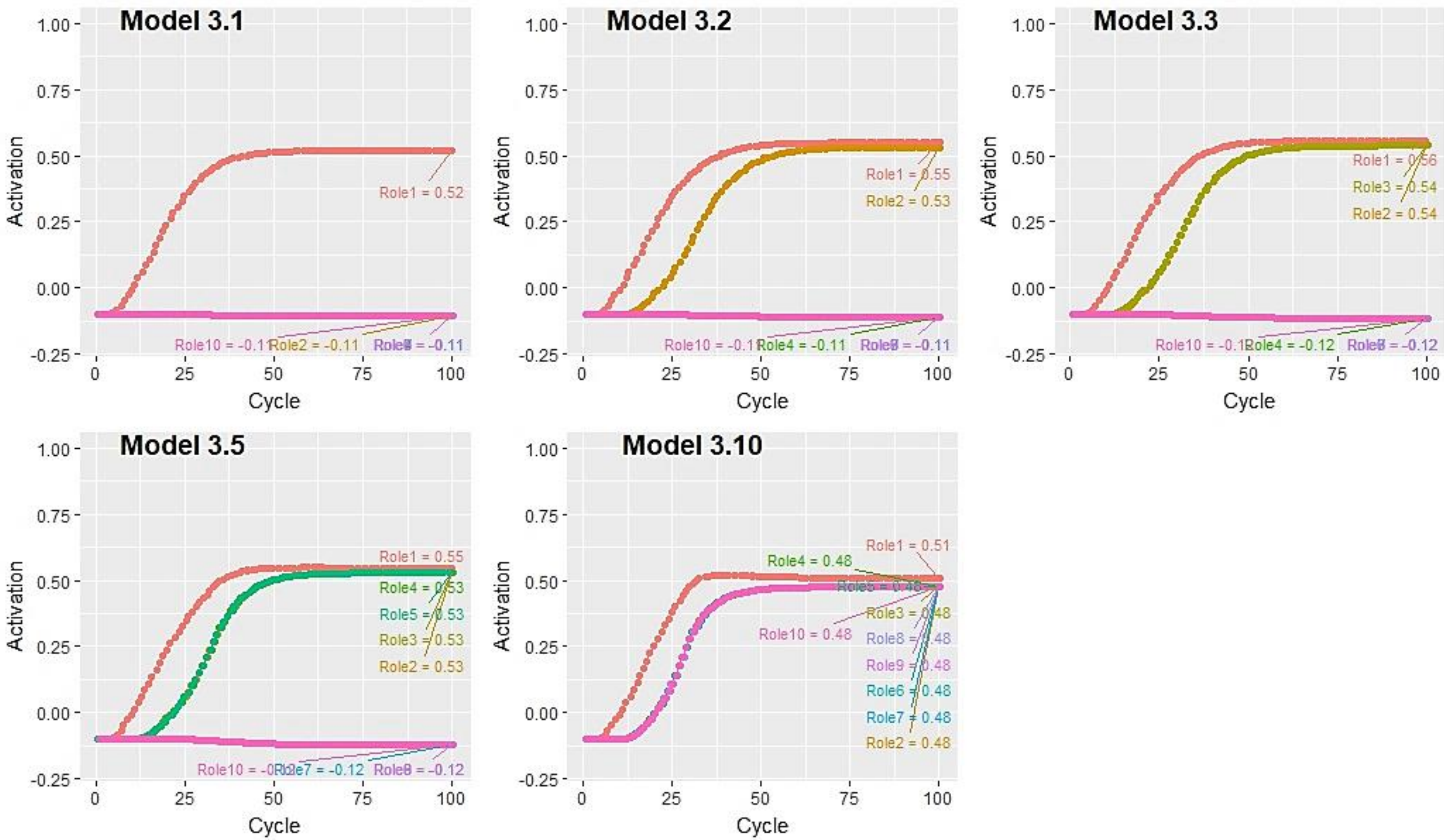

Figure 6.11 Activation levels for Role nodes across the five models after 100 cycles. Panel A refers to the single-perpetrator scenario. The name of each model is listed in each panel (3.1, 3.2, 3.3, 3.5, and 3.10). The input for each scenario was the Perpetrator 1. Role 1 is the role performed by Perpetrator 1, Role 2 is the role performed by Perpetrator 2, and so forth. Recognition is set to the activation level of 0.45.

Differences in Activation Levels Across Models

The differences in the activation levels between PIN 1 and associated PINs, and Role 1 and associated Roles, are listed in Table 6.3. Only the multiple-perpetrator conditions are listed, because no within-pool units were activated for the single-perpetrator condition. The Crimes pool was not included, since only one crime activated in each scenario.

Table 6.3

Activation Levels for Face 1, Person Identity Node Unit 1, Crime 1, and Role 1 Across Five Versions of Model 3

		Activations in PIN Pool			Activations in Role Pool		
Model version	Number of perpetrators	PIN 1	Associated PINs	Difference	Role 1	Associated Roles	Difference
3.2	2	0.64	0.58	0.06	0.55	0.53	0.02
3.3	3	0.64	0.58	0.06	0.55	0.54	0.01
3.5	5	0.64	0.57	0.07	0.54	0.53	0.01
3.10	10	0.61	0.51	0.10	0.50	0.48	0.02

Note. All units in this table reached the recognition threshold of .45. The difference column lists the difference in activation levels between PIN 1 or Role 1, and the associated units within the same pool. For all simulations, external input that modelled Perpetrator 1 was entered into the model.

Between the two types of pools, the difference in activation levels was smaller for Roles than PINs. For example, in the two-perpetrator scenario, PIN 1 and PIN 2 achieve activation levels of 0.64 and 0.58, respectively ($\Delta = 0.06$), but in the same scenario, Role 1 and Role 2 achieve activation levels of 0.55 and 0.53 ($\Delta = 0.02$). Across all four multiple-perpetrator scenarios, the average activation difference between PIN 1 and other PINs was 0.073, and between Role 1 and other Roles was 0.015.

There is no clear indication from the literature how varying degrees of activation should be interpreted. Recognition is treated as a binary condition that occurs at predetermined activation level (0.45 in the current experiment, and in Burton et al., 1990). What remains unclear is whether differing activation levels are accompanied by differential cognitive experiences. For example, does a unit with an activation level of 0.99 elicit the same cognitive response as a unit with an activation level of 0.66? Both units would result in recognition (as they reach the predetermined level of 0.45), but is the first unit accompanied by a sense of confidence that is 50% stronger than the confidence for the second unit?

If the gradient level of activation is accompanied by a gradient level of confidence or memory strength, then the results in Table 6.2. suggest that compared to eyewitness for single-perpetrator crimes, eyewitnesses for multiple-perpetrator crimes will have an overall stronger memory trace for the Crime. Eyewitnesses to multiple-perpetrator crimes will initially have an increased sense of confidence or memory strength for PIN 1 at Set Sizes 2, 3, and 5, compared to eyewitnesses for single-perpetrator crimes. At Set Size 10, the activation level of PIN 1 is less than the activation at Set Size 1. The same trend is seen for the activation levels of Role 1.

In Table 6.3., the differences in activation levels for PIN 1 and associated PINs, and Role 1 and associated Roles are displayed. The difference value for PINs increase as set size increases – this is due to decreased activation for associated pins as set size increases (and a minor decrease in the activation level of PIN 1 at Set Size 10). In contrast, the difference value for Roles hovers around 0.02 and 0.01, but this belies the decrease in activation levels of Role 1 (0.55 to 0.50) and associated Roles (0.53 to 0.48) at Set Size 10.

It is unclear how the activation levels translate into cognitive experiences for the eyewitness. For example, does the overall higher activation level for PINs compared to Roles, imply that eyewitnesses will have a stronger memory for the face of the perpetrator than the role that they performed? If higher activation levels imply greater memory strength, and subsequently, greater confidence, then the increased difference between PIN 1 and associated PINs suggest that eyewitnesses in Model 3.10 (Set Size 10), compared to eyewitnesses in Model 3.2 (Set Size 2), should more easily recognise Perpetrator 1.

If, however, gradient activation levels are associated with gradient cognitive experience, then this may yield an explanation for poor role recollection performance for eyewitnesses to multiple-perpetrator crimes. The difference in activation levels for Role 1 and the associated roles is small (0.01 – 0.02) across all the multiple-perpetrator scenarios. These low activation levels, and subsequent activation differences, may suggest that eyewitnesses to multiple-perpetrator crimes are able to recall all roles, but are unable to differentiate among the roles. Consequently, eyewitnesses to multiple-perpetrator crimes would be able to provide a statement of the crime, but are less likely to accurately delineate perpetrators and roles.

Discussion

The results of the IAC model reported in the current chapter showed that impaired recognition performance associated with increased set sizes is not due to reduced activation of the target face and its associative information. In all the simulations, the activation levels of PIN 1, and its associated units (Role 1 and Crime 1) achieved the recognition threshold. Instead, what the IAC model demonstrated was that in multiple-perpetrator scenarios, the activated PIN for the target (PIN 1) was accompanied by activations of all associated PINs. The activated Role for the target (Role 1) was also accompanied by activations of all associated Roles. The results from the model simulations suggest that (a) perpetrators may be better recognised than roles, and (b) that impaired perpetrator and role recognition is not due to a decrease in memory strength for either the perpetrators or the roles, but is due to difficulty of differentiating among all associated perpetrators and roles instead.

The results of the IAC model are to some extent supported by the results of the three laboratory experiments reported in this thesis, but some of the IAC results are unexplained. The revised IAC model suggests that, overall, memory for roles may be weaker than memory for perpetrators due to the overall higher activation levels for PINS than for Roles.

Compared to memory for perpetrators, the revised IAC model suggests that memory for roles may be more vulnerable to the negative effect of set size. The activation levels of Role 1 decrease as set size increases in the multiple-perpetrator conditions (especially in the Set Size 10 condition), suggesting that role recognition should worsen. The activation levels of the associated Roles are almost the same strength of the activation levels of Role 1 for the multiple-perpetrator conditions, but also decrease as set size increases (especially in Set Size 10). Thus, the IAC results suggest an additive effect: a weakened activation of the role of the perpetrator, accompanied by similar-strength activations of the roles of the associated perpetrators.

There are, however, a few unexpected results. The results of the IAC model suggest that memory for perpetrators may be affected by set size, but these results are in an unexpected direction. The results of the face recognition experiments in Chapter 4 suggest that face recognition performance decreases as set size increases. Instead, the IAC model suggests that the activation level of PIN 1 remains relatively stable across set size. At Set Size 1, the activation level of PIN 1 is 0.62, which increases to 0.64 for Set Sizes 2, 3 and 5, and decreases to 0.61 for Set Size 10 (See Table 6.2). These results suggest that eyewitnesses should be able to recognise Perpetrator 1 across all five Set Size groups. Furthermore, the difference in activation levels between PIN 1 and associated PINs increases as set size increases. If the differential increase in the difference between activation levels is accompanied by an increased confidence or stronger memory trace, then participants should perform better (and report higher confidence) at the recognition task as set size increases. These model results do not mirror the findings from the empirical studies reported in the current thesis.

I am uncertain about how to account for the discrepancy in the results, but I can suggest a few reasons. The IAC model assumes that memory for single-perpetrator and multiple-perpetrator scenarios is encoded at the same strength (all excitatory links are set to 1.0); instead, it is possible that in a multiple-perpetrator scenario memory is weaker compared to memory in a single-perpetrator scenario. If, however, eyewitnesses to multiple-perpetrator crimes were given optimal encoding opportunities (e.g., unlimited time, distinctive targets, clear distinctive roles), then recognition memory may not be as vulnerable to set size as demonstrated in the empirical studies in the current thesis (however, optimal encoding opportunities are highly unlikely).

A counter explanation for the discrepancy in the model results and the empirical results is that the model results may be premature. In the IAC models, the external input was set to Perpetrator 1. This may be analogous to an eyewitness viewing a single-perpetrator parade that contained only Perpetrator 1. If subsequent simulations were run on the multiple-perpetrator versions of the IAC model, where the external input was Perpetrator 2, then Perpetrator 3, and so on, then the results may suggest similar patterns to the semantic priming simulations shown by Burton et al. (1990). The results of the current IAC model may be more akin to comparing recognition results for only the main assailant (in the eyewitness experiment reported in Chapter 5) and first face in the recognition task for the face recognition studies (this is not reported). There is some support for this from the eyewitness experiment reported in Chapter 5: There were no significant differences in recognition accuracy for the main assailant across set size. There was a decrease in Hit Rate across Set Size (although not significant), and there was a significant overall decrease in correct role pairing across Set Size. The negative effect of set size on recognition memory may only be evident in subsequent recognition trials when participants must identify the other perpetrators. Future research should run simulations where different input is shown for each subsequent set of 100 cycles.

Limitations

The divergent results between the IAC model and the laboratory experiments may have emerged from the architecture of the computational models. The excitatory and inhibitory connections for between-pool and within-pool units, as well as their respective weights, are predefined in a set of matrices prior to the simulations. The results of the simulations hinge on these matrices and the architecture of the model (Burton et al., 1990). Thus, the model is heavily dependent on the relations between pools, but also on the relative strengths of these relations.

For the models used in this experiment, all excitatory connections were given the same weight of 1.0, and all inhibitory connections were set to -0.1; these are the same weight used in the IAC built by Burton et al. (1990). Setting all the excitatory connections to the same value (1.0) assumes that the eyewitness encodes all the perpetrators and their roles to an equal extent. Furthermore, setting all the inhibitory connections to the same value (-0.1) assumes that all units exert the same inhibition on one another. Memory for all perpetrators may not be of equal strength – and stronger memory traces should exert more inhibition. Thus, well-encoded memory of a highly distinctive perpetrator may result in stronger excitatory and inhibitory connections. Wells and Pozzulo (2006), for example, found that participants were better able to recognise the accomplice than the main assailant from an identification parade (but there was no difference in role recollection). The difference in recognition accuracy for the accomplice and main assailant could be attributed to idiosyncratic materials, for example, target distinctiveness or the lineup quality. Bindemann et al. (2012), however, provide a counter argument. Their results demonstrate that improved encoding opportunities does not result in improved recognition performance for multiple-perpetrator scenarios. Bindemann et al. reported that participants in the two-target condition still performed worse than participants in the one-target condition even when they were cued about which face to encode for later recognition.

A third limitation of the current IAC model is that it does not 'learn'. As mentioned earlier, the matrices and weights for all connections are predetermined before any simulations are run. This is analogous to only testing recognition, while ignoring the impact of encoding conditions on memory. The relationship between encoding and recognition is important, because these two processes are not independent of each other - instead, their relationship is temporal. Encoding will always precede and influence recognition, and for this reason, optimal recognition conditions cannot compensate for poor encoding. Impaired eyewitness memory for multiple perpetrators may not be a memory problem, but may be evident at perception instead (as suggested by Bindemann et al., 2012; Megreya & Burton, 2006).

The results from the current IAC model suggest that the increased activation levels of all associated perpetrators and roles are due to units within both pools receiving activation from the Crime pool. All associated perpetrators and roles are connected to the same crime unit, and activation of the shared crime unit increases activation input to all connected perpetrators and roles. The Crime pool represents the encoding context where perpetrators and roles are bound together. the current IAC does not test different encoding contexts. Therefore, the current IAC cannot provide any insight into how varying encoding scenarios affect recognition of multiple perpetrators.

A final limitation worth mentioning is the fundamental difference between the IAC model reported by Burton et al. (1990) and the current revised IAC model: The original model represents semantic memory for familiar faces, whereas the revised model represents episodic memory for unfamiliar faces. The architecture and neural representations of episodic and semantic memory may differ, for example, both types of memory may be distributed differently within the brain, have different weights assigned to the respective associative units, and may decay at differential rates depending on the type of memory (e.g., routine episodes may decay faster than semantic memory, whereas unusual, highly distinctive episodes may decay more slowly than semantic memory). The revised IAC model also does not account for when, if, and how memory about a crime changes from episodic memory to semantic memory through the retelling of the event, the natural decay of memory, and the repeated testing of memory through interviews, statements, facial composite construction, and identification parades. One expectation is that memory traces should be strengthened with repeated recollections of the crime, but this may not be uniform for all aspects of the crime (e.g., the memory of the perpetrator's physical appearance versus the narrative of the crime).

Conclusion

The revised IAC model is meant to provide a broad explanation of the difficulties experienced by eyewitnesses to multiple-perpetrator crimes, and it not meant to represent the fine-grained detail of the neural network. Despite the limitations, the revised IAC model attempts to explain how set size affects associative memory and model face recognition and associative memory recognition for unfamiliar faces. This model advances what is known about the relations between faces and associative memory, and how these two types of memory contribute to identity recognition. The IAC model reported in the current chapter also highlights the importance of a shared unit among many other units within the network. The shared unit, which in the current model represents the Crime, suggests that the causal nexus for the relationship between poor recognition performance and set size is the encoding context. All associated roles and perpetrators activate because of being encoded together within the same context, the crime. Future research should test whether accessing other types of associative information, which are unique to the respective perpetrators, will help negate the effect of the shared unit, or whether it will further increase eyewitness memory difficulties.

Chapter 7

General Discussion

The Aims of the Current Thesis

At the beginning of this thesis, I outlined numerous legal challenges surrounding eyewitness memory for multiple-perpetrator crimes. One such challenge is the legal requirement that eyewitness testimony and identification must be sufficiently tested to determine its reliability and to avoid errors such as wrongful convictions. To enforce this requirement at the time of identification, the eyewitness is required to provide supporting information about how they made their lineup decision, which includes the physical appearance of the perpetrator, clothing worn by the perpetrator, or actions performed by the perpetrator during the crime. Providing this type of supporting information for single-perpetrator crimes is easy since any identification made for a single-perpetrator crime simultaneously captures the actions performed by the perpetrator: The perpetrator is recognised as the individual who perpetrated all the actions during the commission of the crime. In contrast, any identification made for a multiple-perpetrator crime does not imply the specific actions that that perpetrator had performed during the crime.

In the case judgements that I reviewed, the judges asserted the belief that eyewitnesses to multiple-perpetrator crimes could differentiate between the actions and identities of the perpetrators who had committed the crimes together. Apart from the judges' belief that eyewitnesses can do this, there are few psychological studies that tested the assertion that eyewitnesses are able to correctly recall each perpetrators' role. Furthermore, I questioned whether eyewitnesses to multiple-perpetrator crimes could identify all the perpetrators who had committed the crime during the lineup procedure.

Although these questions formed the two primary aims of this thesis, they were not the only pivotal questions that needed answering. After reviewing the South African legal texts, it was evident that there were few guidelines that outlined how police officers should administer parades for multiple-perpetrator crimes. In fact, Du Toit et al. (1987) listed only one rule that directly addresses how to administer a parade that contains more than one suspect (Rule 6 in Appendix B): Only up to two suspects may be included in the same parade so long as they share a reasonable level of physical similarity and the parade size is increased. Similar versions of this rule were echoed in Kruger (2017) and the National Instruction (SAPS, 2007). Kruger (2017) further proffered that the official SAPS form, which is used to capture the details of the parade, should also act as a guide for how parades are administered (SAPS 329 form; Appendix A). The SAPS 329 form is misleading in its current form, because the details for up to four suspects are allowed, suggesting that all four suspects can appear together in the same parade. Additionally, South African case law has repeatedly shown that police often place multiple suspects in a single parade together, contributing to the conclusion that parade guidelines are currently not clear.

Recommended and Actual Police Procedure for Administering Parades in South Africa

The paucity of guidance that explicitly addresses lineup administration for multiple suspects and the inconsistency between South legal texts (Du Toit et al., 1987; Kruger, 2017), SAPS guidelines (SAPS, 2007), SAPS 329 form, and case laws led me to question how South African police officers administer multiple-suspect parades,[94] and whether they adapt the recommended lineup procedure.

Evidence exists which demonstrates that police officers adapt recommended procedure when administering parades for multiple suspects: Hobson et al. (2012) reported that police officers in the U.K. adapted both parade instructions and the administration of the parades for eyewitnesses of multiple-perpetrator crimes. However, unlike police forces in the U.K. which administer video parades constructed using video clips from large databases curated for this

[94] See Footnote 1.

purpose, South Africa police officers administer live parades where real people appear behind a one-way mirror so that an eyewitness can make a lineup decision. Due to the type of parade format used, South African police officers may experience different challenges to U.K. officers. Examples of South African case law had partially proved that police officers were adapting lineup procedure, but this had not been investigated in depth.

To investigate how South African officers administer identification parades, I surveyed 75 police detectives who were based at nine different stations in the Western Cape. A notable result was that police officers did not adhere to the recommendations for lineup administration and would often place more than two suspects - and sometimes all suspects - in the same parade. The decision to place more suspects within the same parade than was recommended was motivated by the following reasons: (a) A single parade was quicker and easier to administer than multiple parades; (b) suspects were placed in the same parade if the crime was committed by multiple offenders; and (3) eyewitnesses experienced less distress when viewing a single parade rather than multiple parades. Police officers also reported specific problems with administering live parades, such as finding enough officers to administer the parade, sourcing enough foils to stand in the parade, and the length of time needed to administer the parade.

The results from the survey also supported the legal requirement that eyewitnesses are expected to provide supporting information to justify their lineup identification. While most surveyed police officers reported that eyewitnesses can provide information to support their lineup decisions, officers also stated that eyewitnesses often appear to be confused. Upon prompting, police officers explained that eyewitnesses would confuse the perpetrators with one another or provide information at the time of identification that was new or different from their statement.

One pressing finding from the current survey and review of the South African case law and legal texts is the need to revise current police guidelines so that procedures for multiple-suspect parades are better defined. Although applied psychological research has produced guidelines and recommendations for administering identification parades (for example, Wells & Turtle, 1986; Wells et al., 2000), these guidelines are difficult to implement when investigating multiple-perpetrator crimes. The results of the survey challenged the feasibility of the recommended lineup guidelines when applied from one scenario to another (e.g., single-perpetrator crimes to multiple-perpetrator crimes). Police officers who participated in the study reported numerous difficulties when implementing live parades for multiple-perpetrator crimes. The lack of clear guidelines coupled with the logistical challenges faced by police officers when administering parades result in police officers disregarding recommended line procedure in favour of unconventional techniques; an undesirable outcome when recommended procedures are meant to safeguard against unfair and biased police practices and protect the innocent suspect.

Researchers, legal experts, police officers, and policy makers are faced with an important consideration when developing recommendations: Are the recommendations feasible? There are two questions worth considering when determining the feasibility of solutions to problems surrounding eyewitness memory for multiple-perpetrator crimes. Firstly, will the proposed solution achieve the desired outcomes? For example, will the proposed guideline that each parade be limited to one suspect (a) assist the police with their investigation, (b) aid the eyewitness when making an identification, and (c) protect the innocent suspect from being mistakenly identified? The results from the current thesis suggest that police officers experience multiple difficulties when constructing live parades for multiple-perpetrator crimes. Additionally, asking participants to make identifications from multiple parades does not yield better recognition performance – in fact, eyewitnesses are more likely to adopt a conservative decision-making strategy, which results in them rejecting subsequent parades. The outcome of using multiple parades when investigating multiple-perpetrator crimes is that both innocent *and* guilty suspects are less likely to be identified; a trade-off where recommended police procedure enables the police to judge the reliability of eyewitness memory at the cost of the eyewitness not making any identifications, even if the guilty suspect is present in the parade.

Secondly, can the recommendations be implemented? The legal and psychological recommendation that each identification parade is limited to only one suspect introduces numerous problems when administering live parades: more foils are needed for multiple parades, more police officers are needed to administer the parades, and the entire parade procedure takes longer. Consequently, South African police officers are more likely to adapt the lineup procedure by placing multiple suspects in one parade. What should police officers do when faced with these challenges? The lack of facilities, resources and guidelines does not absolve police officers of the responsibility of using recommended police procedure, but the proposed solutions and guidelines should be applicable to a variety of investigative scenarios, including single-perpetrator and multiple-perpetrator crimes. As is evident from the survey reported in the current thesis and the results reported by Hobson et al. (2012), the current guidelines are not applicable to all investigative scenarios.

One obvious solution to the challenges that police officers experience is to replace live parades with photo or video parades. Even though there is a great deal of psychological research that has investigated whether parade medium has an effect on lineup recognition (e.g., Cutler et al., 1989; Egan et al., 1977; Havard et al., 2010; Kerstholt et al., 2004; Valentine et al., 2007; Valentine et al., 2003; Valentine & Heaton, 1999), and some researchers have argued against using live parades (Fitzgerald, Price, & Valentine, 2018), the South African courts maintain a preference for live parades. It is not clear why South African courts hold this preference; South African legal texts and case law did not offer any insight nor is there any published evidence that South African judges and lawyers believe that live parades are superior to photo or video parades. It would be worthwhile investigating the reason for this preference and whether this preference would change if the courts, judges, and lawyers are presented with convincing evidence. If the courts are more willing to accept other types of parades (e.g., photo parades), then police officers may be more willing to use these types of parades, which would alleviate some of the difficulties that they experience.

Face Recognition Literature and Face Recognition Results

Very practical and obviously important questions about the administration of multiple-suspect parades arose from the results of the survey, but the survey was not able to directly answer the two primary research questions of this thesis: (a) are eyewitnesses to multiple-perpetrator crimes able to identify all the perpetrators, and (b) are eyewitnesses to multiple-perpetrator crimes able to accurately pair perpetrators with their criminal actions? It was possible, however, that these answers could be found in the face recognition and eyewitness memory literatures: The two primary questions address the upper limits of human memory for unfamiliar faces, and how the number of faces at encoding affects face recognition and associative memory (i.e., the ability to correctly recall/remember what information were studied together). A review of the face recognition literature for familiar faces (i.e., faces of individuals whom we know well) provided evidence that familiar faces are robust to the negative effects of time and seeming unbounded; in contrast, the results are less clear for recognition of unfamiliar faces. None of the published manuscripts that explicitly tested the upper limits of face recognition used a similarly high number of items as that used in the recognition studies of non-face images (i.e., upwards of 1000 images). Thus, the maximum number of unfamiliar faces that can be studied and recognised remains unknown, although it is clearly large.

I reviewed three face recognition experiments that manipulated the number of unfamiliar faces that participants studied and tested what effect this manipulation had on subsequent face recognition (Podd, 1990; Lamont et al., 2005; Metzger, 2002). All three studies demonstrated that face recognition performance worsens as the number of faces increased; however, the reason for the impaired performance was equivocal. Podd (1990) reported that recognition worsened because Hits decreased; Lamont et al. (2005) showed that recognition performance worsened due to an increase in False Alarms; whereas Metzger (2002) reported a decrease in Hits and an increase in False Alarms. Unfortunately, neither of these three studies paired any additional information with faces (e.g., names or actions) at encoding, so it was unknown what effect the number of faces shown at encoding would have when testing associative memory.

Only one study manipulated the number of items at encoding and tested both face recognition and associative memory recognition (Bender et al., 2017). Bender et al. (2017) found that the number of items at encoding had no effect on the participants' ability to recognise items that were studied together; however, the authors did not report any descriptive statistics, thus it was not possible to evaluate these results. The non-significant effect of the number of items at encoding on associative memory recognition is surprising, and contrasts with the findings from the face recognition literature that recognition performance decreases as number of items at study increases.

To test the effect of the number of items at study on subsequent recognition of associative memory, I conducted two face recognition experiments in which the number of face-attribute pairs that participants studied was manipulated. After a short delay, participants' memory was tested for (a) the faces that they studied, (b) the attributes that they studied, and (c) which faces and attributes were paired together. The difference between the two face recognition experiments was the type of recognition task used to test participants' ability to pair faces and attributes.

Overall, the results of the two recognition experiments showed that participants performed best at recognising attributes with mixed results for the recognition performance for faces and face-attribute pairs. In the first face recognition experiment, participants performed markedly worse at the pairing task compared to the recognition tests for attributes and faces. For the second face recognition experiment, pairing performance and face recognition performance remained impaired by the number of items that were studied, but participants no longer performed as poorly at the pairing task as participants in the first face recognition experiment.

The results of these two experiments provide support for the negative effect of the number of items at encoding on recognition memory; however, this negative effect is not uniform for all types of memory. While item memory (i.e., memory for faces and attributes) is sensitive to the number of items at encoding, associative memory is more vulnerable.

Although the findings of these two experiments are informative, they did not adequately address the aims of the current thesis. Face recognition research and eyewitness memory research employ different methods at both encoding and recognition, and for this reason, the line of inquiry was extended with an eyewitness memory experiment.

Eyewitness Memory for Multiple-Perpetrators

Within the published eyewitness literature, only 14 published manuscripts were found that investigated eyewitness memory for multiple-perpetrators. Since the methods of the 14 studies differed, it was difficult to compare the results of the draws and draw a definitive conclusion about eyewitness memory for multiple perpetrators. For example, not all studies included a single-perpetrator comparison group. It was clear from the studies that did include a single-perpetrator control group that eyewitness memory was worse in multiple-perpetrator conditions. However, not all the reviewed studies tested recognition of all perpetrators who were involved in the crime. By limiting recognition memory to only one perpetrator, it is not clear whether recognition performance would generalise to all perpetrators – perhaps the poor recognition performance is specific to that perpetrator? Of the few studies in which eyewitnesses were tested on all the perpetrators, recognition results were notably low (ranging between 2% and 14%). Furthermore, only two studies explicitly tested eyewitness memory for the actions performed by the identified perpetrators. Both studies reported that eyewitnesses experienced no difficulties with correctly recalling the roles of identified perpetrators. This result is surprising, because findings from the face recognition literature and the two face recognition experimented that I conducted suggest that associative memory is vulnerable to the negative effects of the number of items at study.

To address the gaps identified within the eyewitness literature, I conducted an eyewitness experiment where I showed participants a video of a crime committed by either one, two, three, five, or ten perpetrators, and tested participants' memory for all the perpetrators involved in the crime. I also tested participants' ability to accurately recall the actions perpetrated by each identified perpetrator during the commission of the crime.

Overall, the results showed that compared to eyewitnesses to single-perpetrator crimes, eyewitnesses to multiple-perpetrator crimes performed worse at the lineup tasks and made fewer correct lineup decisions (i.e., identifying the suspect when he was present, and rejecting the parade when he was not present). In fact, in the high set sizes groups (e.g., Set Size 10), none of the participants correctly recognized all perpetrators. Eyewitnesses to multiple-perpetrator crimes also adopted a stricter criterion, that is, they were more likely to reject the parade than eyewitnesses to single-perpetrator crimes, and this tendency to reject the parade increased as set size increased. These results suggest that for crimes committed by large numbers of perpetrators, eyewitnesses are unlikely to make accurate decisions for all parades that they view and may reject subsequent parades instead.

The findings of the conducted eyewitness experiment support the results reported in the eyewitness literature: Eyewitnesses to multiple-perpetrator crimes perform worse at the lineup task than eyewitnesses to single-perpetrator crimes. This experiment also provides insight into the questions that arose when reviewing the eyewitness literature for multiple-perpetrator crimes, for example, was eyewitness recognition for multiple perpetrators impaired overall or only for the one perpetrator for whom eyewitnesses were tested? I isolated the results for the main assailant in my experiment and compared them across the various experimental groups; there were no significant differences between groups. When I compared overall lineup performance, however, it was obvious that recognition performance decreased as the number of perpetrators increased – in fact, no one who encoded a crime committed by 10 perpetrators achieved 100% accuracy across all the parades. One implication of these results is that real-world eyewitnesses are less likely to correctly identify guilty suspects if they are present in the parade or correctly reject the parade if they are not.

Implications of Role Recollection for Multiple-Perpetrator Crimes

One novel finding of this experiment is that compared to eyewitnesses to single-perpetrator crimes, eyewitnesses to multiple-perpetrator crimes are worse at role recollection. The results suggest a trend where the likelihood of correctly recalling the actions perpetrated by each perpetrator decreases as the number of perpetrators increase. This finding is especially notable when considering that role recollection can only be correct if it follows a correct identification. In the current experiment, however, a correct lineup identification did not guarantee correct role recollection: Participants who identified the correct perpetrator did not necessarily recall the correct role!

Poor role recollection may have serious implications for real-world investigations: Would the courts accept the testimony and identifications made by seemingly reliable eyewitnesses to multiple-perpetrator crimes if aspects of their testimony and identification are incongruent with the information provided in their statement? Although the police, lawyers, and judges have no way of measuring the accuracy of an eyewitness' statement and identification, they may be presented with an eyewitness who provides contradictory evidence at the time of their identification compared to information in the statement (or even compared to other eyewitnesses of the same crime); in such situations, would the identification and testimony still be considered reliable? Would such an eyewitness be asked to testify in court? Or does the incorrect role recollection have no bearing on the investigation, which continues forward regardless? The answers to these questions are unknown. Future research should consider surveying judges and lawyers on this question to find out their legal opinion of such a scenario, and whether the testimony and identification of eyewitnesses to multiple-perpetrator crimes would be dismissed, have diminished value or have no effect in such a situation.

Strict Criterion: More Questions than Answers

The strict criterion, impaired identification performance, and impaired role recollection of eyewitnesses to multiple-perpetrator crimes has significant bearing on current police procedure. The strict criterion adopted for multiple–perpetrator crimes presents the police with a practical problem. Administering live parades is resource-expensive: Police officers must arrange multiple foils to appear in the parade, multiple police officers to assist with the parade, there are transport costs and security risks, and administering parades takes time. It is not desirable to hold an identification parade if there is a high risk of the eyewitness rejecting the parade, and it is even less desirable to do so if the parade includes the guilty suspect.

In the current thesis, I decided that participants would view one parade after another, and after making an identification from each parade, would move to the next parade. Although my decision was motivated by current recommended police procedure in South Africa, the results of the eyewitness study pose a new question: Is the strict criterion transient, that is, will the inclination to reject the parade disappear if a longer interval is introduced between parades? Some effects within the applied psychological literature are transient; for example, the verbal overshadowing effect (Schooler & Engstler-Schooler, 1990) reduces when a delay is introduced between the description task and identification task (Clare & Lewandowsky, 2004).

Furthermore, it is unknown whether the criterion shift is induced by the number of parades viewed consecutively, or the number of targets seen at study. Some evidence exists in the eyewitness literature where eyewitness' choosing behaviour is affected by the number of parades viewed following encoding of a single-perpetrator crime. For example, when participants are shown two parades - first, a blank parade that is known to contain only innocent foils and no suspects, and a subsequent parade that may or may not contain a suspect - participants appear to adopt a stricter criterion for the second parade (see the results in both Palmer, Brewer, & Weber, 2012 and Wells, 1984). Neither Palmer et al. (2012) nor Wells (1984) formally analysed the rejection rates, but both studies suggest that criterion (i.e., choosing behaviour) changes because of multiple parades.

Criterion may, however, be dependent on the interval between parades. If administering parades on the same day has a negative, knock-on effect on subsequent parades, then police officers could consider increasing the interval between parades. Longer intervals may yield more promising results, although at a greater logistical cost since the eyewitness would have to return to the police station at a later stage.

Lineup Formats: Experimental Method Versus Real-World Practice

There is one obvious difference between the lineup task used in my eyewitness experiment and the lineup procedure that South African police officers reported to use: Participants in the eyewitness experiment saw multiple parades, which were limited to only one suspect (or no suspects), whereas actual police officers reported that they would often place all the suspects in the same parade together. It is unknown whether the recognition performance will be the same for both parade types. The method often used by South African police officers – where all suspects are placed in a large parade together – is similar to the mixed-suspect parade that Wells and Turtle (1986) discouraged. Wells and Turtle (1986) argued that a single-suspect parade was a better investigative tool than an all-suspect parade (i.e., a parade that only contains suspects and no foils) and mixed-suspect parade (i.e., a parade that contains more than one suspect and some foils). The mixed-suspect parade that Wells and Turtle argued against differs from the mixed-suspect parade implemented by South African police: In the first scenario, there is one perpetrator who committed the crime but multiple suspects (of which only one can be guilty), whereas in the second scenario, there are multiple perpetrators who committed the crime and multiple suspects - each representing one perpetrator.

None of the empirical studies that I reviewed compared recognition performance for a mixed-suspect parade to recognition performance for multiple single-suspect parades within a single experiment. Furthermore, I am unaware of any empirical evidence that clearly demonstrates that eyewitnesses to multiple-perpetrator crimes are more confused when viewing a mixed-suspect parade than multiple single-suspect parades. Besides the suggestion from Wells and Turtle (1986) that mixed-suspect parades are less desirable than single-suspect parades, is there any empirical evidence that supports this hypothesis for parades for multiple-perpetrator crimes? This suggestion may seem controversial, but we should question *why* it is controversial: The single-suspect parade, which is adopted into the canon of recommended police procedure, was specifically recommended for single-perpetrator crimes. Currently there is no empirical evidence that has demonstrated that eyewitnesses to multiple-perpetrators perform worse at a mixed-suspect parade than multiple single-suspect parades (although it is unclear how large a mixed-suspect parade must be).

Furthermore, a single, large parade may induce a criterion that is less strict than that induced by multiple, single-suspect parade. If the eyewitness' choosing behaviour is affected by the number of parades that they view, then viewing only one parade may prevent the eyewitness from changing their criterion for subsequent parades. If, however, the strict criterion is due to the number of perpetrators who committed the crime, rather than the number of lineups, then eyewitnesses to multiple-perpetrator crimes should adopt a strict choosing behaviour when they view the first parade. The results of the experiment reported in Chapter 5 do not suggest that rejection rates for the main assailant (which was always the first parade) increased as set size increased. Future research should replicate the Bayesian analysis reported by Wells and Turtle (1986) for a multiple-perpetrator scenario, and test the hypothesis that multiple single-suspect parades are superior to a single mixed-suspect parade in the laboratory.

What is to be Done About Role Recollection?

Regardless of which parade format police officers implement, they remain faced with the possibility that eyewitnesses to multiple-perpetrator crimes cannot accurately recall the actions performed by the identified perpetrator. As discussed in Chapter 1 and Chapter 5, role recollection for each perpetrator has serious implications for the investigation and for sentencing. In the reported experiments, I tested only identification and role recollection following identification; further research is needed to determine whether eyewitnesses to multiple-perpetrator crimes experience the same degree of impairment when recollecting the criminal event, like when giving a statement. Some evidence exists to suggest that both recollection of the criminal event and description of the perpetrators are negatively impacted by the number of perpetrators who committed the crime. Fahsing et al. (2004) reported that the descriptions of the perpetrators are more impoverished when given by eyewitnesses to multiple-perpetrator crimes than eyewitnesses to single-perpetrator crimes. These results support the findings of Clifford and Hollin (1981): Participants in the violent-crime condition yielded descriptions of the perpetrator that decreased in quality as the number of perpetrators increased.

While these results suggest that recollection memory is impaired by the number of perpetrators, it is possible that recollection can be bolstered through an innovative interviewing technique. If recollection and recognition are differentially affected by the number of perpetrators at encoding, then there may be methods to improve recollection of the multiple-perpetrator crimes and the descriptions of the perpetrators. For example, the eye-closure interview (ECI) leads to better recall of a criminal event by reducing cognitive load (Vredeveldt, Hitch, & Baddeley, 2011). Since recollection of multiple-perpetrator crimes is likely associated with a higher cognitive load than recollection of single-perpetrator crimes, the ECI may be particularly helpful for multiple-perpetrator crimes. Another possibility is to test innovative lineup instructions to help eyewitnesses better recall the roles performed by the perpetrators. Hobson and Wilcock (2011) found that using a simple instruction, such as asking eyewitnesses to recall the role of the crime while viewing the parade, resulted in improved role recollection after viewing the parade. A combination of the ECI and an adapted lineup instruction may help eyewitnesses to recall the roles at the time of identification. assist eyewitness better recall the roles of each perpetrator: For example, an innovative interview technique, like the ECI, may lead to successful recollection of the roles performed by each of the perpetrators, and this may strengthen the memory traces between each perpetrator and the actions that they performed during the crime. At the time of identification, a revised lineup instruction, like the one used by Hobson and Wilcock, may help eyewitnesses access the roles performed by the perpetrators, especially if these memory traces were strengthened by the ECI. Future research should consider investigating whether there are ways to improve role recollection for eyewitnesses, and whether improved role recollection results in better lineup identification and subsequent role pairing.

Computational Models and Insight from the IAC Model

Following the three experiments reported in the current thesis, it was unclear why eyewitnesses to multiple-perpetrator crimes struggled to recall the roles associated with the identified perpetrators. Previous research had used connectionist neural networks to explain how the priming of semantic information resulted in faster reaction times for primed versus non-primed faces. Of all the face recognition models reviewed, only two types – connectionist neural networks and functional models – included both associative information and face recognition within the model architecture. Furthermore, the Interactive Activation and Competition (IAC) network has been used to model different scenarios for face recognition (e.g., Burton et al., 1990).

Since the IAC was reliably used in previous research investigation various face recognition scenarios and included both associative memory and faces, I adapted this neural network to explain the recognition results and role recollection impairments for multiple-perpetrator crimes. There were two notable changes to the structure underpinning the IAC network used in previous experiments: First, the revised version simulated recognition of unfamiliar faces; second, the revised version included both semantic and episodic information, rather than only semantic information.

The overlap between the constructs, episodic memory and semantic memory, was particularly interesting, and it remains unclear to me how semantic memory and episodic memory are formed, what the nature of their relationship is, and how they influence each other. The way a single external event/experience leads to the creation of both episodic and semantic memory is especially interesting, as well as how these two types of memory for the same event are represented cognitively. The distinction between episodic and semantic memory is important because Burton et al. (1990) originally included semantic information in their model for familiar face recognition and semantic priming. For familiar faces, which are encoded through multiple events, the episodic event in which the semantic information is first learnt either fades or is diluted when the semantic information is reinforced across multiple events. In contrast, eyewitnesses to crimes do not have multiple opportunities to encode the perpetrator when the perpetrator is a stranger. If the architecture of the original IAC model was retained to model eyewitness memory for perpetrators and their actions, then it seems bizarre that an eyewitness can recall the actions of the perpetrator (i.e., semantic information) without also recalling the episodic content of the crime. However, if semantic memory develops from episodic memory, which can fade (i.e., is forgotten) or is reinforced across multiple episodes (i.e., so that one single episode is more important than another) then episodic and semantic memory should exist separately. A network architecture where episodic memory and semantic memory exist independently would allow for the recollection of semantic memory without the simultaneous recollection of episodic memory. For this reason, I treated the crime as an episodic event, which existed separately from the roles committed by the perpetrators.

The primary result from the revised IAC model was that all associated information encoded together in the same episodic context will activate when any part of that information is activated. These results suggest that when eyewitnesses to multiple-perpetrator crimes view an identification parade that contains any one of the guilty suspects, then all associated information for that crime activates - that is, the actions committed by all the perpetrators and the memories of the physical appearance of all the perpetrators activate, and consequently, the eyewitness is unable to differentiate among the activated information. It is for this reason that eyewitnesses to multiple-perpetrator crimes may be able to correctly recall the crime and the actions perpetrated by the perpetrators of the crime when making a statement but will confuse the actions between the perpetrators. The activation of competing information accounts for why eyewitnesses struggle to recognise the perpetrator and differentiate among the roles committed within the crime.

The findings from the IAC simulations are interesting but warrant further investigation. The IAC model makes several assumptions, for example, that all targets are encoded at the same strength, that the amount of cognitive resources needed to suppress other competing information is in finite, and that roles are encoded at equal strength; these assumptions are very unlikely in a real-world scenario. The study of memory is simultaneously a study of encoding: Optimal recognition conditions cannot compensate for poor encoding. Since the IAC model does not take encoding into account, it treats the multiple-perpetrator disadvantage as a memory problem. Some research, however, has shown that viewing two faces, rather than one face, leads to impaired matching and delayed matching performance (Bindemann et al., 2012; Megreya & Burton, 2006); these results challenge the view that the multiple-perpetrator disadvantage is a memory problem since the disadvantage is already present at encoding and perception.

The IAC model does not provide a full explanation of the eyewitness disadvantage associated with multiple-perpetrator crimes, but it does contribute to the field of computational psychological models and highlights some interesting challenges in construct definition. Future research should extend the IAC model to include different parade format types, because the current model assumes that only one suspect is shown at a time. An interesting extension of the current revised IAC model would be to replicate the semantic priming modelled by Burton et al. (1990) but adapted for the criminal context. That is, would viewing one perpetrator prime recognition of another perpetrator?

Looking Forward

This thesis started with two questions: Can eyewitnesses to multiple-perpetrator crimes identify all the perpetrators who committed the crime? Can eyewitnesses to multiple-perpetrator crimes correctly recall the actions performed by each perpetrator? I employed a multi-faceted approach to answer these questions: a police survey, two face recognition experiments, an eyewitness experiment, and an artificial neural network. Instead of producing only two simple answers, this thesis yielded a myriad of exciting research questions that warrant further investigation.

This thesis can offer the following conclusions: Compared to eyewitnesses for single-perpetrator crimes, eyewitnesses to multiple-perpetrator crimes will most likely perform poorly across multiple lineup tasks and are less likely to accurately recall the roles performed by the perpetrators. Furthermore, performance for both tasks worsens as the number of perpetrators increases. Recognising that eyewitnesses to multiple-perpetrator crimes experience recognition memory and pairing difficulties is extremely important for various reasons. Firstly, current police procedure for administering lineups must be adapted appropriately. Secondly, police officers, judges, and lawyers must adapt their expectations of the memory capabilities of eyewitnesses to multiple-perpetrator crimes, who may not be able to accurately recall the role of each perpetrator. Thirdly, in the absence of other corroborating evidence, the courts should consider adopting the principle of common purpose for all multiple-perpetrator crimes where the roles of the perpetrators cannot be reliably disentangled by the eyewitness. Since eyewitnesses to multiple-perpetrator crimes experience unique challenges when viewing an identification parade, and subsequent role recollection, a uniform legal and investigative approach to these witnesses – as has been adopted up to this point – is not appropriate. With the current prevalence of multiple-perpetrator crimes in South Africa, it is imperative that key stakeholders take heed of the memory difficulties that eyewitnesses to multiple-perpetrator crimes experience, as well as the shortcomings of current recommended guidelines for administering multiple-suspect parades; only then can appropriate and feasible procedures be developed to test eyewitness memory for multiple-perpetrator crimes.

References

Anderson, R. C., & Pichert, J. W. (1978). Recall of previously unrecallable information following a shift in perspective. *Journal of Verbal Learning and Verbal Behavior*, *17*(1), 1-12. doi:10.1016/s0022-5371(78)90485-1

Arkes, H. R., & Freedman, M. R. (1984). A demonstration of the costs and benefits of expertise in recognition memory. *Memory & Cognition*, *12*(1), 84-89. doi:10.3758/bf03197000

Atkinson, R. C., & Juola, J. F. (1974). Search and decision processes in recognition memory. In D. H. Krantz, R. C. Atkinson, R. D. Luce, & P. Suppes (Eds.), Contemporary developments in mathematical psychology: I. Learning, memory and thinking. Oxford, England: W. H. Freeman.

Australian Bureau of Statistics. (2004). Sexual Assault in Australia: a statistical overview. Canberra: Australian Bureau of Statistics. Retrieved from http://www.ausstats.abs.gov.au/ausstats/subscriber.nsf/0/C41F8B2864D42333CA256F070079CBD4/$File/45230_2004.pdf

Bäckman, L. (1991). Recognition memory across the adult life span: The role of prior knowledge. *Memory & Cognition*, *19*(1), 63-71. doi:10.3758/bf03198496

Bäckman, L., & Herlitz, A. (1990). The relationship between prior knowledge and face recognition memory in normal aging and Alzheimer's disease. *Journal of Gerontology*, *45*(3), 94-100. doi:10.1093/geronj/45.3.p94

Baddeley, A. (2001). The concept of episodic memory. *Philosophical Transactions of the Royal Society of London*. *Series B, 356*(1413), 1345–1350. doi:10.1098/rstb.2001.0957

Bahrick, H. P., Bahrick, P. O., & Wittlinger, R. P. (1975). Fifty years of memory for names and faces: A cross-sectional approach. *Journal of Experimental Psychology: General*, *104*(1), 54–75. doi:10.1037//0096-3445.104.1.54

Bakdash, J. Z., & Marusich, L. R. (2017). Repeated measures correlation. *Frontiers in Psychology*, *8*, 1-13. doi:10.3389/fpsyg.2017.00456

Barry, C., Johnston, R. A., & Scanlan, L. C. (1998). Are Faces “Special” Objects? Associative and Sem antic Priming of Face and Object Recognition and Naming. *The Quarterly Journal of Experimental Psychology Section A, 51*(4), 853-882. doi:10.1080/713755783

Barton, K. (2018). MuMIn: Multi-Model Inference. R package version 1.40.4. https://CRAN.R-project.org/package=MuMIn

Bastin, C., & Van der Linden, M. (2005). The effects of aging on the recognition of different types of associations. *Experimental Aging Research, 32*(1), 61-77. doi:10.1080/03610730500326291

Bastin, C., Van der Linden, M., Schnakers, C., Montaldi, D., & Mayes, A. R. (2010). The contribution of familiarity to within-and between-domain associative recognition memory: Use of a modified remember/know procedure. *European Journal of Cognitive Psychology, 22*(6), 922-943. doi:10.1080/09541440903007834

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 1-48. doi:10.18637/jss.v067.i01

Becker, C. A. (1979). Semantic context and word frequency effects in visual word recognition. *Journal of Experimental Psychology: Human Perception and Performance, 5*(2), 252–259. doi:10.1037//0096-1523.5.2.252

Bender, A. R., Naveh-Benjamin, M., Amann, K., & Raz, N. (2017). The role of stimulus complexity and salience in memory for face–name associations in healthy adults: Friend or foe?. *Psychology and Aging, 32*(5), 489–505. doi:10.1037/pag0000185

Bindemann, M., Sandford, A., Gillatt, K., Avetisyan, M., & Megreya, A. M. (2012). Recognising faces seen alone or with others: Why are two heads worse than one?. *Perception, 41*(4), 415-435. doi:10.1068/p6922

Bower, G. H. (2000). A brief history of memory research. In E. Tulving & F. I. M. Craik (Eds.), *The Oxford handbook of memory*, (pp. 3-32).New York: Oxford University Press.

Brady, T. F., Konkle, T., Alvarez, G. A., & Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences, 105*(38), 14325-14329. doi:10.1073/pnas.0803390105

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

Bruce, V. (1982). Changing faces: Visual and non-visual coding processes in face recognition. *British Journal of Psychology, 73*(1), 105-116. doi:10.1111/j.2044-8295.1982.tb01795.x

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

Bruce, V., & Young, A. (1986). Understanding face recognition. *British Journal of Psychology, 77*(3), 305-327. doi:10.1111/j.2044-8295.1986.tb02199.x

Burton, A. M. (1994). Learning new faces in an interactive activation and competition model. *Visual Cognition, 1*(2-3), 313-348. doi:10.1080/13506289408402304

Burton, A. M., & Bruce, V. (1992). I recognize your face but I can't remember your name: A simple explanation?. *British Journal of Psychology, 83*(1), 45-60. doi:10.1111/j.2044-8295.1992.tb02424.x

Burton, A. M., & Bruce, V. (1993). Naming faces and naming names: Exploring an interactive activation model of person recognition. *Memory, 1*(4), 457-480. doi:10.1080/09658219308258248

Burton, A. M., Bruce, V., & Johnston, R. A. (1990). Understanding face recognition with an interactive activation model. *British Journal of Psychology, 81*(3), 361-380. doi:10.1111/j.2044-8295.1990.tb02367.x

Clare, J., & Lewandowsky, S. (2004). Verbalizing facial memory: criterion effects in verbal overshadowing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(4), 739–755. doi:10.1037/0278-7393.30.4.739

Clark, S. E. (2003). A memory and decision model for eyewitness identification. *Applied Cognitive Psychology, 17*(6), 629-654. doi:10.1002/acp.891

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

Clark, S. E., & Tunnicliff, J. L. (2001). Selecting lineup foils in eyewitness identification experiments: Experimental control and real-world simulation. *Law and Human Behavior, 25*(3), 199–216. doi:10.1023/a:1010753809988

Clifford, B. R., & Hollin, C. R. (1981). Effects of the type of incident and the number of perpetrators on eyewitness memory. *Journal of Applied Psychology, 66*(3), 364–370. doi:10.1037//0021-9010.66.3.364

Coates, T. (1999). The strange story of Adolph Beck. London: Stationery Office Press.

Cousineau, D. (2005). Confidence intervals in within-subject designs: A simpler solution to Loftus and Masson’s method. *Tutorials in Quantitative Methods for Psychology, 1*(1), 42-45. doi:10.20982/tqmp.01.1.p042

Craik, F. I., Govoni, R., Naveh-Benjamin, M., & Anderson, N. D. (1996). The effects of divided attention on encoding and retrieval processes in human memory. *Journal of Experimental Psychology: General*, *125*(2), 159-180. doi:10.1037//0096-3445.125.2.159

Crook, T. H., III, Larrabee, G. J., & Youngjohn, J. R. (1993). Age and incidental recall for a simulated everyday memory task. *Journal of Gerontology, 48*(1), 45-47. doi:10.1093/geronj/48.1.p45

Crottaz-Herbette, S., Anagnoson, R. T., & Menon, V. (2004). Modality effects in verbal working memory: differential prefrontal and parietal responses to auditory and visual stimuli. *Neuroimage, 21*(1), 340-351. doi:10.1016/j.neuroimage.2003.09.019

Curran, K., & Millie, A. (2003). Rape and indecent assault: Incidence and service provision in Southwark. *London: Safer Southwark Partnership.* Retrieved from https://dspace.lboro.ac.uk/dspace-jspui/bitstream/2134/943/1/Curran%20and%20Millie%202003%20-%20report[1].pdf

Cutler, B. L., Fisher, R. P., & Chicvara, C. L. (1989). Eyewitness identification from live versus videotaped lineups. *American Journal of Forensic Psychology, 26*(4), 67-81.

Cutler, B. L., Penrod, S. D., & Martens, T. K. (1987). The reliability of eyewitness identification: The role of system and estimator variables. *Law and Human Behavior*, *11*(3), 233-258. doi:10.1007/bf01044644

Da Silva, T., Woodhams, J., & Harkins, L. (2013). Heterogeneity within multiple perpetrator rapes: A national comparison of lone, duo, and 3+ perpetrator rapes. *Sexual Abuse: A Journal of Research and Treatment*, *26*(6), 503-522. doi:10.1177/1079063213497805

Davies, G., & Griffiths, L. (2008). Eyewitness identification and the English courts: A century of trial and error. *Psychiatry, Psychology and Law, 15*(3), 435-449. doi:10.1080/13218710802101605

DeCarlo, L. T. (2012). On a signal detection approach to m-alternative forced choice with bias, with maximum likelihood and Bayesian approaches to estimation. *Journal of Mathematical Psychology, 56*(3), 196-207. doi:10.1016/j.jmp.2012.02.004

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

Dempsey, J. L., & Pozzulo, J. D. (2008). Identification accuracy of eyewitnesses for a multiple perpetrator crime: Examining the simultaneous and elimination lineup procedures. *American Journal of Forensic Psychology, 26*(4), 67-81.

Dempsey, J. L., & Pozzulo, J. D. (2012). Children's identification accuracy of multiple perpetrators: Examining the simultaneous versus elimination line-up. *Psychiatry, Psychology and Law, 20*(3), 353-365. doi:10.1080/13218719.2012.679124

Devlin, Lord P. 1976. Report to the Secretary of State for the Home Department on the Departmental Committee on Evidence of Identification in Criminal Cases, London: HMSO.

Diamond, R., & Carey, S. (1986). Why faces are and are not special: an effect of expertise. *Journal of Experimental Psychology: General, 115*(2), 107-117. doi:10.1037//0096-3445.115.2.107

Dobolyi, D. G., & Dodson, C. S. (2013). Eyewitness confidence in simultaneous and sequential lineups: A criterion shift account for sequential mistaken identification overconfidence. *Journal of Experimental Psychology: Applied, 19*(4), 345-357. doi:10.1037/a0034596

Du Toit, E., de Jager, F. J., Paizes, A., Skeen, A., & van der Merwe, S. E. (1987). Commentary on the Criminal Procedure Act. Claremont: Juta and Company (PTY) LTD.

Egan, D., Pittner, M., & Goldstein, A. G. (1977). Eyewitness identification: Photographs vs. live models. *Law and Human Behavior, 1*(2), 199-206. doi:10.1007/bf01053439

Ellis, H. D., Davies, G. M., & Shepherd, J. W. (1978). A critical examination of the Photofit system for recalling faces. *Ergonomics, 21*(4), 297-307. doi:10.1080/00140137808931726

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

European Union Agency for Fundamental Rights (FRA). (2012). European Union Minorities and Discrimination Survey (EU-MIDIS). Data in Focus Report: Minorities as Victims of Crime. Belgium: European Union Agency for Fundamental Rights.

Fahsing, I. A., Ask, K., & Granhag, P. A. (2004). The man behind the mask: accuracy and predictors of eyewitness offender descriptions. *Journal of Applied Psychology, 89*(4), 722-729. doi:10.1037/0021-9010.89.4.722

Farah, M. J. (1996). Is face recognition ‘special’? Evidence from neuropsychology. *Behavioural Brain Research, 76*(1-2), 181-189. doi:10.1016/0166-4328(95)00198-0

Farah, M. J., Wilson, K. D., Drain, M., & Tanaka, J. N. (1998). What is" special" about face perception?. *Psychological Review, 105*(3), 482-498. doi:10.1037//0033-295x.105.3.482

Fitzgerald, R., Price, H. L., & Valentine, T. (2018). Eyewitness identification: live, photo, and video lineups. *Psychology, Public Policy, and Law*.

Fox, J., & Weisberg, S. (2011). An {R} Companion to Applied Regression, Second Edition. Thousand Oaks CA: Sage.

Franklin, K. (2004). Enacting masculinity: Antigay violence and group rape as participatory theater. *Sexuality Research & Social Policy*, *1*(2), 25-40. doi:10.1525/srsp.2004.1.2.25

Gardiner, J. M., & Java, R. I. (1990). Recollective experience in word and nonword recognition. *Memory & Cognition*, *18*(1), 23-30. doi:10.3758/bf03202642

Godfrey, R. D., & Clark, S. E. (2010). Repeated eyewitness identification procedures: Memory, decision making, and probative value. *Law and Human Behavior*, *34*(3), 241-258. doi:10.1007/s10979-009-9187-7

Goldstein, A. G. (1977). The fallibility of the eyewitness: Psychological evidence. In B. D. Sales (Ed.), Psychology in the legal process (pp. 223 – 248). New York: Spectrum Publications, Inc.

Goldstein, A. G., & Mackenberg, E. J. (1966). Recognition of human faces from isolated facial features: A developmental study. *Psychonomic Science*, *6*(4), 149-150. doi:10.3758/bf03328001

Gronlund, S. D., Wixted, J. T., & Mickes, L. (2014). Evaluating eyewitness identification procedures using receiver operating characteristic analysis. *Current Directions in Psychological Science*, *23*(1), 3-10. doi:10.1177/0963721413498891

Gunter, B., Furnham, A. F., & Gietson, G. (1984). Memory for the news as a function of the channel of communication. *Human Learning: Journal of Practical Research & Applications*, *3*(4), 265-271.

Haig, N. D. (1984). The effect of feature displacement on face recognition. *Perception*, *13* (5), 505-512. doi:10.1068/p130505

Hancock, P. J., Burton, A. M., & Bruce, V. (1996). Face processing: Human perception and principal components analysis. *Memory & Cognition*, *24*(1), 26-40. doi:10.3758/bf03197270

Havard, C., Memon, A., Clifford, B., & Gabbert, F. (2010). A comparison of video and static photo lineups with child and adolescent witnesses. *Applied Cognitive Psychology*, *24*(9), 1209-1221. doi:10.1002/acp.1645

Hay, D. C., & Young, A.W., & Ellis, A. W. (1991). Routes through the face recognition system. *The Quarterly Journal of Experimental Psychology*, 43A (4), 761-791. https://doi.org/10.1080/14640749108400957

Hobson, Z. J. (2013). Eyewitness identification of multiple perpetrators. (Unpublished Doctorate of Philosophy thesis). London South Bank University, England.

Hobson, Z. J., & Wilcock, R. (2011). Eyewitness identification of multiple perpetrators. *International Journal of Police Science & Management, 13*(4), 286-296. doi:10.1350/ijps.2011.13.4.253

Hobson, Z., Wilcock, R., & Valentine, T. (2012). Multiple Suspect Showing: A Survey of Police Identification Officers. *Policing: A Journal of Policy and Practice, 7*(1), 79-87. doi:10.1093/police/pas021

Home Office of the United Kingdom. (2017). CODE D Revised: Code of Practice for the identification of persons by Police Officers. Police and Criminal Evidence Act 1984 (PACE). Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/592562/pace-code-d-2017.pdf

Horvath, M. A. H., & Kelly, L. (2009). Multiple perpetrator rape: Naming an offence and initial research findings. *Journal of Sexual Aggression, 15*(1), 83-96. doi:10.1080/13552600802653818

Howells, T. H. (1938). A study of ability to recognize faces. *The Journal of Abnormal and Social Psychology, 33*(1), 124–127. doi:10.1037/h0062942

Ihlebæk, C., Løve, T., Eilertsen, D. E., & Magnussen, S. (2003). Memory for a staged criminal event witnessed live and on video. *Memory, 11*(3), 319-327. doi:10.1080/09658210244000018

Jacob, P. (1994). The feasibility of using multiple-suspect sequential lineups. (Unpublished Bacheor of Arts [Honours] thesis). Queen’s University, Ontario.

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language, 30*(5), 513-541. doi:10.1016/0749-596x(91)90025-f

James, L. E., Fogler, K. A., & Tauber, S. K. (2008). Recognition memory measures yield disproportionate effects of aging on learning face-name associations. *Psychology and Aging, 23*(3), 657-664. doi:10.1037/a0013008

Jang, Y., Wixted, J. T., & Huber, D. E. (2009). Testing signal-detection models of yes/no and two-alternative forced-choice recognition memory. *Journal of Experimental Psychology: General, 138*(2), 291-306. doi:10.1037/a0015525

Jewkes, R., Vetten, L., Jina, R., Christofides, N., Sigsworth, R., & Loots, L. (2012). What we know-and what we don't: Single and multiple perpetrator rape in South Africa. South African Crime Quarterly, 41, 11-19. doi:10.17159/2413-3108/2012/v0i41a835

Johnson, P. C. D. (2014). Extension of Nakagawa & Schielzeth's R2 GLMM to random slopes models. *Methods in Ecology and Evolution. 5*(9), 944–946. doi:10.1111/2041-210x.12225

Johnston, R. A., & Edmonds, A. J. (2009). Familiar and unfamiliar face recognition: A review. *Memory, 17*(5), 577-596. doi:10.1080/09658210902976969

Joubert, S. J., Naudé, C. M. B., Neser, J. J., van der Hoven, A. E., Snyman, W. J., & van der Merwe, E. (2002). Introduction to Criminology. Module 2. Causation, Explanation, Prevention and Reaction to Crime. Pretoria: University of South Africa.

Kanwisher, N. (2000). Domain specificity in face perception. *Nature Neuroscience, 3*(8), 759-763. doi:10.1038/77664

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience, 17*(11), 4302-4311. doi:10.1523/jneurosci.17-11-04302.1997

Kaufmann, J. M., & Schweinberger, S. R. (2004). Expression influences the recognition of familiar faces. *Perception, 33*(4), 399-408. doi:10.1068/p5083

Kerstholt, J. H., Koster, E. R., & van Amelsvoort, A. G. (2004). Eyewitnesses: A comparison of live, video, and photo line-ups. *Journal of Police and Criminal Psychology, 19*(2), 15-22. doi:10.1007/bf02813869

Konkle, T., Brady, T. F., Alvarez, G. A., & Oliva, A. (2010a). Conceptual distinctiveness supports detailed visual long-term memory for real-world objects. *Journal of Experimental Psychology: General, 139*(3), 558-578. doi:10.1037/a0019165

Konkle, T., Brady, T. F., Alvarez, G. A., & Oliva, A. (2010b). Scene memory is more detailed than you think: The role of categories in visual long-term memory. *Psychological Science, 21*(11), 1551-1556. doi:10.1177/0956797610385359

Krause, B. J., Schmidt, D., Mottaghy, F. M., Taylor, J., Halsband, U., Herzog, H., ... & Müller-Gärtner, H. W. (1999). Episodic retrieval activates the precuneus irrespective of the imagery content of word pair associates: a PET study. *Brain, 122*(2), 255-263. doi:10.1093/brain/122.2.255

Kruger, A. (2017). Hiemstra Criminal Procedure. South Africa: LexisNexis South Africa.

Laldin, S. (1997). Contextual effect on lineup identification of multiple perpetrators. (Unpublished Bacheor of Science [Honours] thesis). Queen's University, Ontario.

Lamont, A. C., Stewart-Williams, S., & Podd, J. (2005). Face recognition and aging: Effects of target age and memory load. *Memory & Cognition, 33*(6), 1017-1024. doi:10.3758/bf03193209

Lee, M. D., & Wagenmakers, E. J. (2014). Bayesian cognitive modeling: A practical course. Cambridge university press.

Lindsay, R. C., Pozzulo, J. D., Craig, W., Lee, K., & Corber, S. (1997). Simultaneous lineups, sequential lineups, and showups: Eyewitness identification decisions of adults and children. *Law and Human Behavior, 21*(4), 391-404. doi:10.1023/a:1024807202926

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

Liu, C. H., & Chaudhuri, A. (2003). What determines whether faces are special?. *Visual Cognition, 10*(4), 385-408. doi:10.1080/13506280244000050

Loftus, E. F. (1976). Unconscious transference in eyewitness identification. *Law & Psychological Review, 2*, 93-98.

Loftus, G. R., & Masson, M. E. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin & Review*, *1*(4), 476-490. doi:10.3758/bf03210951

Lowndes, G. J., Saling, M. M., Ames, D., Chiu, E., Gonzalez, L. M., & Savage, G. (2008). Recall and recognition measures of paired associate learning in healthy aging. *Aging, Neuropsychology, and Cognition, 15*(4), 506-522. doi:10.1080/13825580802099678

Luus, C. A., & Wells, G. L. (1991). Eyewitness identification and the selection of distracters for lineups. *Law and Human Behavior, 15*(1), 43-57. doi:10.1007/bf01044829

Macmillan, N. A. & Creelman, C. D. (2005). Detection Theory: A User's Guide. 2nd edition. New Jersey: Lawrence Erlbaum Associates.

Malpass, R. S., & Kravitz, J. (1969). Recognition for faces of own and other race. *Journal of Personality and Social Psychology, 13*(4), 330-334. doi:10.1037/h0028434

Malpass, R. S., Tredoux, C. G., & McQuiston-Surrett, D. (2007). Lineup construction and lineup fairness. In R.C. L. Lindsay, D. E. Ross, J. D. Read, & M. P. Toglia (Eds), *The handbook of eyewitness psychology, Vol II: Memory for people* (pp. 155-178). New Jersey: Lawrence Erlbaum Mahwah.

Mandler, G. (1991). Your face looks familiar but I can't remember your name: A review of dual process theory. In W.E. Murdock & S. Lewandowsky (Eds.), *Relating theory and data: Essays on human memory in honor of Bennet B. Murdock* (pp. 207-225). New Jersey: Lawrence Erlbaum Associates, Inc.

Maw, S. (2012). The psychological impact of rape trauma: A longitudinal study of adult female survivors in the Western Cape, South Africa. (Unpublished doctoral thesis). University of Cape Town, Cape Town.

Mayes, A., Montaldi, D., & Migo, E. (2007). Associative memory and the medial temporal lobes. *Trends in Cognitive Sciences, 11*(3), 126-135. doi:10.1016/j.tics.2006.12.003

McClelland, J. L. (2015). Explorations in Parallel Distributed Processing: A Handbook of Models, programs and exercises. Retrieved from https://web.stanford.edu/group/pdplab/pdphandbook/.

McClelland, J. L., Rumelhart, D. E., & the PDP Research Group (1986). Parallel distributed processing: explorations in the microstructure of cognition. Volume 1. Foundations. MIT Press, Cambridge MA.

McKelvie, S. J. (1976). The role of eyes and mouth in the memory of a face. *The American Journal of Psychology, 89* (2), 311-323. doi:10.2307/1421414

Mecklenburg, S. H. (2006). Report to the legislature of the state of illinois: The illinois pilot program on double-blind, sequential lineup procedures. http://www.chicagopolice.org/IL%20Pilot%20on%20Eyewitness%20ID.pdf. (Main Report), http://www.chicagopolice.org/Apndx%20to%20IL%20Pilot%20on%20Eyewitness%20ID.pdf (Appendix), and http://eyewitness.utep.edu/Documents/IllinoisPilotStudyOnEyewitnessIDAddendum.pdf (Addendum).

Megreya, A. M., & Bindemann, M. (2011). Identification accuracy for single-and double-perpetrator crimes: Does accomplice gender matter?. *British Journal of Psychology, 103*(4), 439-453. doi:10.1111/j.2044-8295.2011.02084.x

Megreya, A. M., & Burton, A. M. (2006). Recognising faces seen alone or with others: When two heads are worse than one. *Applied Cognitive Psychology, 20*(7), 957-972. doi:10.1002/acp.1243

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

Meissner, C. A., Tredoux, C. G., Parker, J. F., & MacLin, O. H. (2005). Eyewitness decisions in simultaneous and sequential lineups: A dual-process signal detection theory analysis. *Memory & Cognition, 33*(5), 783-792. doi:10.3758/bf03193074

Metzger, M. M. (2002). Stimulus load and age effects in face recognition: A comparison of children and adults. *North American Journal of Psychology, 4*(1), 51-61.

Meyer, R. (2014). Computes activation according to the _interactive activation and competition_. CompANN package. Retrieved from https://github.com/ReneMayer/compANN/blob/master/DESCRIPTION

Mickes, L., Flowe, H. D., & Wixted, J. T. (2012). Receiver operating characteristic analysis of eyewitness memory: Comparing the diagnostic accuracy of simultaneous versus sequential lineups. *Journal of Experimental Psychology: Applied*, *18*(4), 361-376. doi:10.1037/a0030609

Morey, R. D. (2008). Confidence intervals from normalized data: A correction to Cousineau (2005). *Tutorial in Quantitative Methods for Psychology*, *4*(2), 61-64. doi:10.20982/tqmp.04.2.p061

Morton, J. (1969). Interaction of information in word recognition. *Psychological Review, 76*(2), 165-178. doi:10.1037/h0027366

Moscovitch, M., Cabeza, R., Winocur, G., & Nadel, L. (2016). Episodic memory and beyond: the hippocampus and neocortex in transformation. *Annual Review of Psychology, 67*(1), 105-134. doi:10.1146/annurev-psych-113011-143733

Moscovitch, M., Winocur, G., & Behrmann, M. (1997). What is special about face recognition? Nineteen experiments on a person with visual object agnosia and dyslexia but normal face recognition. *Journal of Cognitive Neuroscience, 9*(5), 555-604. doi:10.1162/jocn.1997.9.5.555

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution, 4* (2), 133–142. doi:10.1111/j.2041-210x.2012.00261.x

Naveh-Benjamin, M. (2000). Adult age differences in memory performance: tests of an associative deficit hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*(5), 1170-1187. doi:10.1037//0278-7393.26.5.1170

Naveh-Benjamin, M., Craik, F. I., Guez, J., & Dori, H. (1998). Effects of divided attention on encoding and retrieval processes in human memory: Further support for an asymmetry. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*(5), 1091-1104. doi:10.1037//0278-7393.24.5.1091

Naveh-Benjamin, M., Guez, J., Kilb, A., & Reedy, S. (2004). The associative memory deficit of older adults: Further support using face-name associations. *Psychology and Aging, 19*(3), 541-546. doi:10.1037/0882-7974.19.3.541

Naveh-Benjamin, M., Shing, Y. L., Kilb, A., Werkle-Bergner, M., Lindenberger, U., & Li, S. C. (2009). Adult age differences in memory for name–face associations: The effects of intentional and incidental learning. *Memory, 17*(2), 220-232. doi:10.1080/09658210802222183

Navon, D. (1992). Selection of lineup foils by similarity to the suspect is likely to misfire. *Law and Human Behavior, 16*(5), 575-593. doi:10.1007/bf01044624

Nickerson, R. S. (1965). Short-term memory for complex meaningful visual configurations: A demonstration of capacity. Canadian Journal of Psychology/Revue canadienne de psychologie, 19(2), 155-160. doi:10.1037/h0082899

Nickerson, R. S. (1968). A note on long-term recognition memory for pictorial material. *Psychonomic Science, 11*(2), 58-58. doi:10.3758/bf03330991

Old, S. R., & Naveh-Benjamin, M. (2008). Memory for people and their actions: Further evidence for an age-related associative deficit. *Psychology and Aging, 23*(2), 467-472. doi:10.1037/0882-7974.23.2.467

Oliva, A., & Torralba, A. (2007). The role of context in object recognition. *Trends in Cognitive Sciences, 11*(12), 520-527. doi:10.1016/j.tics.2007.09.009

Owens, J. N. (2009). Is three really a crowd? The effects of multiple perpetrators on eyewitness identification accuracy. (Unpublished Masters of Arts [Forensic Psychology] thesis). John Jay College of Criminal Justice of the City University of New York, United States.

Perea, M., & Rosa, E. (2002). The effects of associative and semantic priming in the lexical decision task. *Psychological Research, 66*(3), 180-194. doi:10.1007/s00426-002-0086-5

Phillips, P. J., Moon, H., Rizvi, S. A., & Rauss, P. J. (2000). The FERET evaluation methodology for face-recognition algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22*(10), 1090-1104. doi:10.1109/34.879790

Podd, J. (1990). The effects of memory load and delay on facial recognition. *Applied Cognitive Psychology, 4*(1), 47-59. doi:10.1002/acp.2350040105

Pozzulo, J. D., & Warren, K. L. (2003). Descriptions and identifications of strangers by youth and adult eyewitnesses. *Journal of Applied Psychology, 88*(2), 315-323. doi:10.1037/0021-9010.88.2.315

Putnam, A. L., Ozubko, J. D., MacLeod, C. M., & Roediger, H. L. (2014). The production effect in paired-associate learning: benefits for item and associative information. *Memory & Cognition, 42*(3), 409-420. doi:10.3758/s13421-013-0374-x

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Rakover, S. S., & Cahlon, B. (2001). Face recognition: Cognitive and computational processes (Vol. 31). Amsterdam, The Netherlands: John Benjamins Publishing.

Reinitz, M. T., & Demb, J. B. (1994). Implicit and explicit memory for compound words. *Memory & Cognition, 22*(6), 687-694. doi:10.3758/bf03209253

Reinitz, M. T., & Hannigan, S. L. (2001). Effects of simultaneous stimulus presentation and attention switching on memory conjunction errors. *Journal of Memory and Language, 44*(2), 206-219. doi:10.1006/jmla.2000.2727

Rhodes, M. G., & Anastasi, J. S. (2012). The own-age bias in face recognition: a meta-analytic and theoretical review. *Psychological Bulletin, 138*(1), 146. http://dx.doi.org/10.1037/a0025750

Rhodes, M. G., Castel, A. D., & Jacoby, L. L. (2008). Associative recognition of face pairs by younger and older adults: the role of familiarity-based processing. *Psychology and Aging, 23*(2), 239-249. doi:10.1037/0882-7974.23.2.239

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J. C., Müller, N. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics, 12* (1), 77-84. doi:10.1186/1471-2105-12-77

Robinson, M. D., Johnson, J. T., & Herndon, F. (1997). Reaction time and assessments of cognitive effort as predictors of eyewitness memory accuracy and confidence. *Journal of Applied Psychology, 82*(3), 416-425. doi:10.1037//0021-9010.82.3.416

Russo, R., Ward, G., Geurts, H., & Scheres, A. (1999). When unfamiliarity matters: Changing environmental context between study and test affects recognition memory for unfamiliar stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*(2), 488-499. doi:10.1037//0278-7393.25.2.488

Rust, A., & Tredoux, C. (1998). Identification parades: An empirical survey of legal recommendations and police practice in South Africa. *South African Journal of Criminal Justice, 11* (2), 196-212. Retrieved from https://heinonline-org.ezproxy.uct.ac.za/HOL/Page?handle=hein.journals/soafcrimj11&div=25&start_page=196&collection=journals&set_as_cursor=0&men_tab=srchresults

Sadr, J., Jarudi, I., & Sinha, P. (2003). The role of eyebrows in face recognition. *Perception, 32*(3), 285-293. doi:10.1068/p5027

Schiff, W., Banka, L., & de Bordes Galdi, G. (1986). Recognizing people seen in events via dynamic" mug shots". *The American Journal of Psychology, 99*(2), 219-231. doi:10.2307/1422276

Schooler, J. W., & Engstler-Schooler, T. Y. (1990). Verbal overshadowing of visual memories: Some things are better left unsaid. *Cognitive Psychology, 22*(1), 36–71. doi:10.1016/0010-0285(90)90003-m

Shapiro, P. N., & Penrod, S. (1986). Meta-analysis of facial identification studies. *Psychological Bulletin, 100*(2), 139-156. doi:10.1037//0033-2909.100.2.139

Shaw, J. S., III (1996). Increases in eyewitness confidence resulting from postevent questioning. *Journal of Experimental Psychology: Applied, 2*(2), 126-146. doi:10.1037//1076-898x.2.2.126

Shaw, J. S., III, & McClure, K. A. (1996). Repeated postevent questioning can lead to elevated levels of eyewitness confidence. *Law and Human Behavior, 20*(6), 629-653. doi:10.1007/bf01499235

Sheehan, C. (2011). Making the jurors the" experts": The case for eyewitness identification jury instructions. Boston College. Law School. Boston College Law Review, 52(2), 651. Retrieved from https://lawdigitalcommons.bc.edu/cgi/viewcontent.cgi?article=3159&context=bclr

Shepard, R. N. (1967). Recognition memory for words, sentences, and pictures. *Journal of Verbal Learning and Verbal Behavior, 6*(1), 156-163. doi:10.1016/s0022-5371(67)80067-7

Shepherd, J. W. (1983). Identification after long delays. In S. M. A. Lloyd-Bostock & B.R. Clifford (Eds.) Evaluating Witness Evidence: Recent Psychological Research and new perspectives (pp. 173 – 187). Great Britain: John Wiley & Sons Ltd.

Sourcebook of Criminal Justice Statistics. (2008). Estimated percent distribution of violent victimizations by multiple offenders by type of crime and perceived race of offenders, Retrieved from http://www.albany.edu/sourcebook/pdf/t3312008.pdf.

South African Police Service. (2007). National Instruction: identification parades. Pretoria: Commissioner of the SAPS

South African Police Service. (2014). Crime Situation in South Africa (Released 29 September 2015). Retrieved from https://www.saps.gov.za/services/crimestats_archive.php

South African Police Service. (2017). Crime Situation in South Africa: 1 April 2016 – 31 March 2017. Retrieved from https://www.saps.gov.za/services/final_crime_stats_presentation_24_october_2017.pdf

Sporer, S. L. (2001a). Recognizing faces of other ethnic groups: An integration of theories. *Psychology, Public Policy, and Law, 7*(1), 36-97. doi:10.1037//1076-8971.7.1.36

Sporer, S. L. (2001b). The cross-race effect: Beyond recognition of faces in the laboratory. *Psychology, Public Policy, and Law, 7*(1), 170-200. doi:10.1037//1076-8971.7.1.170

Sporer, S. L., Penrod, S., Read, D., & Cutler, B. (1995). Choosing, confidence, and accuracy: A meta-analysis of the confidence-accuracy relation in eyewitness identification studies. *Psychological Bulletin, 118*(3), 315-327. doi:10.1037//0033-2909.118.3.315

Standing, L. (1973). Learning 10000 pictures. *Quarterly Journal of Experimental Psychology, 25*(2), 207-222. doi:10.1080/14640747308400340

Standing, L., Conezio, J., & Haber, R. N. (1970). Perception and memory for pictures: Single-trial learning of 2500 visual stimuli. *Psychonomic Science, 19*(2), 73-74. doi:10.3758/bf03337426

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers, 31*(1), 137-149. doi:10.3758/bf03207704

Statistics South Africa (2017). Victims of Crime Survey 2016-2017. Pretoria: Statistics South Africa. Data retrieved from https://www.datafirst.uct.ac.za/dataportal/index.php/catalog/511

Statistics South Africa (2014). Victims of Crime Survey 2013-2014. Pretoria: Statistics South Africa. Retrieved from https://www.statssa.gov.za/publications/P0341/P03412016.pdf

Steblay, N. M. (1992). A meta-analytic review of the weapon focus effect. *Law and Human Behavior, 16*(4), 413-424. doi:10.1007/bf02352267

Swart, L. A., Gilchrist, A., Butchart, A., Seedat, M., & Martin, L. (2000). Rape surveillance through district surgeon offices in Johannesburg, 1996–1998: findings, evaluation and prevention implications. *South African Journal of Psychology, 30*(2), 1-10. doi:10.1177/008124630003000201

Tarr, M. J., & Gauthier, I. (2000). FFA: a flexible fusiform area for subordinate-level visual processing automatized by expertise. *Nature Neuroscience, 3*(8), 764-769. doi:10.1038/77666

Tredoux, C., Nunez, D., Oxtoby, O., & Prag, B. (2006). An evaluation of ID: an eigenface based construction system: reviewed article. *South African Computer Journal, 2006*(37), 90-97. Retrieved from https://www.researchgate.net/profile/Oliver_Oxtoby/publication/228903986_An_evaluation_of_ID_An_eigenface_based_construction_system/links/54d878c60cf2970e4e77c7cf/An-evaluation-of-ID-An-eigenface-based-construction-system.pdf

Tsunga, L. (2018). The post-identification feedback effect and eyewitness choosing behaviour. (Unpublished Master's thesis). University of Cape Town, South Africa.

Tulving, E. (1972). Episodic and semantic memory. Organization of memory. In E. Tulving & W. Donaldson, *Organization of memory (pp. 381-403).* Oxford, England: Academic Press.

Tulving, E. (1985a). How many memory systems are there?. *American Psychologist, 40*(4), 385-398. doi:10.1037//0003-066x.40.4.385

Tulving, E. (1985b). Memory and consciousness. *Canadian Psychology/Psychologie canadienne, 26*(1), 1-12. doi:10.1037/h0080017

Tunnicliff, J. L., & Clark, S. E. (2000). Selecting foils for identification lineups: Matching suspects or descriptions?. *Law and Human Behavior, 24*(2), 231-258. doi:10.1023/a:1005463020252

Tupper, N. (2018). Eyewitness identification for multiple perpetrator crimes: Examine recognition memory and decisions for multiple faces. (Unpublished Doctorate of Philosophy thesis). University of Portsmouth, England, and Masstricht University, The Netherlands.

Turk, M., & Pentland, A. (1991). Eigenfaces for recognition. *Journal of Cognitive Neuroscience, 3*(1), 71-86. doi:10.1162/jocn.1991.3.1.71

Valentine, T. (1991). A unified account of the effects of distinctiveness, inversion, and race in face recognition. *The Quarterly Journal of Experimental Psychology Section A, 43*(2), 161-204. doi:10.1080/14640749108400966

Valentine, T., Darling, S., & Memon, A. (2007). Do strict rules and moving images increase the reliability of sequential identification procedures?. *Applied Cognitive Psychology, 21*(7), 933-949. doi:10.1002/acp.1306

Valentine, T., Harris, N., Piera, A. C., & Darling, S. (2003). Are police video identifications fair to African–Caribbean suspects?. *Applied Cognitive Psychology, 17*(4), 459-476. doi:10.1002/acp.880

Valentine, T., & Heaton, P. (1999). An evaluation of the fairness of police lineups and video identifications. *Applied Cognitive Psychology, 13*(S1), S59-S72.doi: 10.1002/(SICI)1099-0720(199911)13:1+<S59::AID-ACP679>3.0.CO;2-Y

Vanderwal, A. (1996). The effects of the sequential presentation of lineups with multiple perpetrators on eyewitness identification. (Unpublished Bacheor of Arts [Honours] thesis). Queen's University, Ontario.

Vogt, S., & Magnussen, S. (2007). Long-term memory for 400 pictures on a common theme. *Experimental Psychology, 54*(4), 298-303. doi:10.1027/1618-3169.54.4.298

Vredeveldt, A., Hitch, G. J., & Baddeley, A. D. (2011). Eyeclosure helps memory by reducing cognitive load and enhancing visualisation. Memory & cognition, 39(7), 1253-1263. https://doi.org/10.3758/s13421-011-0098-8

Vredeveldt, A., Tredoux, C. G., Nortje, A., Kempen, K., Puljević, C., & Labuschagne, G. N. (2015). A field evaluation of the Eye-Closure Interview with witnesses of serious crimes. *Law and Human Behavior, 39*(2), 189-197. doi:10.1037/lhb0000113

Wall, P. M. (1965). Eye-witness identification in criminal cases. United States of America: Charles C. Thomas.

Watkins, M. J., Ho, E., & Tulving, E. (1976). Context effects in recognition memory for faces. *Journal of Verbal Learning and Verbal Behavior, 15*(5), 505-517. doi:10.1016/0022-5371(76)90045-1

Wells, G. L. (1978). Applied eyewitness-testimony research: System variables and estimator variables. *Journal of Personality and Social Psychology, 36*(12), 1546-1557. doi:10.1037//0022-3514.36.12.1546

Wells, G. L. (1984). The Psychology of Lineup Identifications 1. *Journal of Applied Social Psychology, 14*(2), 89-103. doi:10.1111/j.1559-1816.1984.tb02223.x

Wells, G. L. (2006). Eyewitness identification: Systemic reforms. *Wisconsin Law Review*, 615-643. Retrieved from https://www.researchgate.net/profile/Gary_Wells3/publication/265280996_Eyewitness_identification_Systemic_reforms/links/54871bc10cf2ef34478eb3a3/Eyewitness-identification-Systemic-reforms.pdf

Wells, G. L., & Bradfield, A. L. (1998). " Good, you identified the suspect": Feedback to eyewitnesses distorts their reports of the witnessing experience. *Journal of Applied Psychology, 83*(3), 360-376. doi:10.1037//0021-9010.83.3.360

Wells, G. L., Leippe, M. R., & Ostrom, T. M. (1979). Guidelines for empirically assessing the fairness of a lineup. *Law and Human Behavior, 3*(4), 285-293. doi:10.1007/bf01039807

Wells, G. L., Malpass, R. S., Lindsay, R. C. L., Fisher, R. P., Turtle, J. W., & Fulero, S. M. (2000). From the lab to the police station: A successful application of eyewitness research. *American Psychologist, 55*(6), 581-598. doi:10.1037//0003-066x.55.6.581

Wells, E. C., & Pozzulo, J. D. (2006). Accuracy of eyewitnesses with a two-culprit crime: Testing a new identification procedure. *Psychology, Crime & Law, 12*(4), 417-427. doi:10.1080/10683160500050666

Wells, G. L., Rydell, S. M., & Seelau, E. P. (1993). The selection of distractors for eyewitness lineups. *Journal of Applied Psychology, 78*(5), 835-844. doi:10.1037//0021-9010.78.5.835

Wells, G. L., Small, M., Penrod, S., Malpass, R. S., Fulero, S. M., & Brimacombe, C. E. (1998). Eyewitness identification procedures: Recommendations for lineups and photospreads. *Law and Human Behavior, 22*(6), 603-647. doi:10.1023/a:1025750605807

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

Whiffen, J. W., & Karpicke, J. D. (2017). The role of episodic context in retrieval practice effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 43*(7), 1036-1046. doi:10.1037/xlm0000379

Wilcock, R. A. & Kneller, W. (2011). A comparison of presentation methods of video identification parades. *Applied Cognitive Psychology, 25*(6), 835-840. https://doi.org/10.1002/acp.1754

Winograd, E., & Rivers-Bulkeley, N. T. (1977). Effects of changing context on remembering faces. *Journal of Experimental Psychology: Human Learning and Memory, 3*(4), 397-405. doi:10.1037//0278-7393.3.4.397

Wittwer, T. (2018). The own-group bias in face recognition: improving accuracy with training. (Unpublished doctoral thesis). University of Cape Town, Cape Town and Université Toulouse II - Jean Jaurès, Toulouse, France.

Wixted, J. T., Mickes, L., Clark, S. E., Gronlund, S. D., & Roediger, H. L., III(2015). Initial eyewitness confidence reliably predicts eyewitness identification accuracy. *American Psychologist, 70*(6), 515-526. doi:10.1037/a0039510

Wogalter, M. S., Malpass, R. S., & Mcquiston, D. E. (2004). A national survey of US police on preparation and conduct of identification lineups. *Psychology, Crime and Law, 10*(1), 69-82. doi:10.1080/10683160410001641873

Yarmey, A. D. (1982). Eyewitness identification and stereotypes of criminals. In A.Trankell (Ed.) Reconstructing the past: The role of psychologists in criminal trials (pp. 205-225).Deventer, The Netherlands: Kluwer.

Yonelinas, A. P. (1994). Receiver-operating characteristics in recognition memory: evidence for a dual-process model. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(6), 1341-1354. doi:10.1037//0278-7393.20.6.1341

Yonelinas, A. P. (2001). Consciousness, control, and confidence: the 3 Cs of recognition memory. *Journal of Experimental Psychology: General*, *130*(3), 361-379. doi:10.1037//0096-3445.130.3.361

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language, 46*(3), 441-517. doi:10.1006/jmla.2002.2864

Young, A. W., Hay, D. C., & Ellis, A. W. (1985). The faces that launched a thousand slips: Everyday difficulties and errors in recognizing people. *British Journal of Psychology, 76*(4), 495-523. doi:10.1111/j.2044-8295.1985.tb01972.x

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

Yonelinas, A. P., & Parks, C. M. (2007). Receiver operating characteristics (ROCs) in recognition memory: a review. *Psychological Bulletin, 133*(5), 800-832. doi:10.1037/0033-2909.133.5.800

Young, A. W., Hellawell, D., & Hay, D. C. (1987). Configurational Information in Face Perception. *Perception, 16*(6), 747-759. doi:10.1068/p160747

Yue, X., Tjan, B. S., & Biederman, I. (2006). What makes faces special?. *Vision Research, 46*(22), 3802-3811. doi:10.1016/j.visres.2006.06.017

Appendix A

SAPS 329 form

SOUTH AFRICAN POLICE SERVICE — SUID-AFRIKAANSE POLISIEDIENS

IDENTIFICATION PARADE FORM — **UITKENNINGSPARADEVORM**

Section 37 (1) (b) of Act No. 51 of 1977 — Artikel 37 (1) (b) van Wet No. 51 van 1977

1. Member in charge of parade
Lid in beheer van parade..

2. Station
Stasie .... CAS/CR No.
MAS/MR No........../........../..........

3. Charge(s)/Klagte(s)..
....
....
....

4. Instruction for the institution of the parade received on
Opdrag vir die hou van die parade ontvang op..
at
om.. from the investigating officer No.
vanaf die ondersoekbeampte No ....
Rank
Rang.. Name
Naam ....

5. Full name(s) of suspect(s):
Volle naam(e) van verdagte(s):
(1) .... Language
Taal..
(2) .... Language
Taal..
(3) .... Language
Taal..
(4).. Language
Taal..
(If there are more suspects, continue on folio)/(In geval van meer verdagtes, vervolg op folio)

6. Suspect(s) was/were informed on
Verdagte(s) is op.. of the intended
van die voorgenome
parade on
parade op.. at
om.. at
te.. ingelig.

7. Suspect(s) was/were informed on
Verdagte(s) is op.. that he/she/they
verwittig dat hy/sy/hulle
is/are entitled to legal representation.
geregtig is op regsverteenwoordiging.

8. Suspect(s) desire/do not desire legal representation.
Verdagte(s) verlang/verlang nie regsverteenwoordiging/nie.

9. Legal representative(s):
Regsverteenwoordiger(s): (1).. (2) ....
(3).. (4).. was/were informed
is
on
op (1).. (2)..
(3) ....(4) ....
of the date, time and place of the parade.
ingelig van die datum, tyd en plek van die parade.

10. Name of photographer
Naam van fotograaf..

11. Name of interpreter
Naam van tolk..

12. The parade was held out of sight and hearing of other persons.
Die parade is buite sig en gehoor van ander persone gehou.
Place
Plek.. date
datum.. time
tyd ....

13. Name of member who guarded the witness(es) before he/she/they attended the parade
Naam van lid wat toesig oor getuie(s) gehou het voordat hy/sy/hulle parade bygewoon het..
.... Office No.
Kantoor No..

14. Name of member who escorts witness(es) **to** the parade
Naam van lid wat getuie(s) **na** parade begelei..
....

15. Name of member who escorts witness(es) **from** the parade
Naam van lid wat getuie(s) **van** parade begelei..
....

16. Name of member who guarded the witness(es) after he/she/they had attended the parade
Naam van lid wat toesig oor getuie(s) gehou het nadat hy/sy/hulle parade bygewoon het..
.... Office No.
Kantoor No..

17. There were
Daar was altesaam.. persons on parade [including the suspect(s)]. They are of about the same height, build,
persone op parade [insluitende die verdagte(s)]. Hulle is min of meer van dieselfde lengte,
age and appearance andwere dressed more or less the similar to the suspect(s)
liggaamsbou, ouderdom en voorkoms en is almal naastenby soos die verdagte(s) geklee.

18. Suspect(s) has been informed of the allegation(s) and the purpose of the parade and that he/she/they—
Verdagte(s) is van die beweerde klagte(s) en die doel van die parade verwittig en meegedeel dat hy/sy/hulle—

(1) may take up any position of his/her/their choice on the parade and may change his/her/their position before another witness is called; and
enige posisie van sy/haar/hulle keuse op die parade mag inneem en van posisie mag verander voordat 'n ander getuie geroep word; en

(2) may make any reasonable request(s) in respect of the parade.
enige redelike versoek(e) ten opsigte van die parade mag rig.

19. (1) His/her/their request(s) is/are the following:
Sy/haar/hulle versoek(e) is soos volg:

Suspect/Verdagte 1 ....

....

Suspect/Verdagte 2 ....

....

Suspect/Verdagte 3 ....

....

Suspect/Verdagte 4 ....

(2) Steps taken as a result of the request(s):
Stappe gedoen as gevolg van die versoek(e):

Suspect/Verdagte 1 ....

....

Suspect/Verdagte 2 ....

....

Suspect/Verdagte 3 ....

....

Suspect/Verdagte 4 ....

20. Suspect(s) was/were asked whether he/she/they is/are satisfied with the parade, including the persons on parade.
Verdagte(s) is gevra of hy/sy/hulle tevrede is met die opstel van die parade, insluitende die persone op parade.
His/her/their answer(s) is/are as follows:
Sy/haar/hulle antwoord(e) is soos volg:

Suspect/Verdagte 1 ....

....

Suspect/Verdagte 2 ....

....

Suspect/Verdagte 3 ....

....

Suspect/Verdagte 4 ....

....

21. Name of legal representative(s), if present:
Naam van regsverteenwoordiger(s), indien teenwoordig:

Name/Naam	On behalf of/Namens
....	....
....	....
....	....
....	....

22. Persons on parade, suspect(s) included:
Persone op parade, verdagte(s) ingesluit:

Name/Naam	Age/Ouderdom	Address/Adres
(1)..............................		
(2)..............................		
(3)..............................		
(4)..............................		
(5)..............................		
(6)..............................		
(7)..............................		
(8)..............................		
(9)..............................		
(10)..............................		
(11)..............................		

Name/Naam	Age/Ouderdom	Address/Adres
(12)........................	....	....
(13)........................	....	....
(14)........................	....	....
(15)........................	....	....
(16)........................	....	....
(17)........................	....	....
(18)........................	....	....
(19)........................	....	....
(20)........................	....	....

23. A photograph was taken after the parade had been set up: Yes/No
'n Foto is geneem nadat die parade opgestel is: Ja/Nee

PROCEDURE OF PARADE/VERLOOP VAN PARADE

24. The persons, including the suspect(s), taking part in the parade, take up the following positions from left to right in front of me:
Die persone, insluitende die verdagte(s), wat aan die parade deelneem, neem die volgende posisies van links na regs voor my in:

(1) .... (2)........................
(3) .... (4)........................
(5) .... (6)........................
(7) .... (8)........................
(9) .... (10)........................
(11) .... (12)........................
(13) .... (14)........................
(15) .... (16)........................
(17) .... (18)........................
(19) .... (20)........................

25. ***First witness/Eerste getuie:***

Name/Naam .... Language/Taal ....
was asked to point out the suspect(s), if on parade, by touching his/her/their shoulder(s), who (date, time, place and charge):
is gevra om die verdagte(s), indien op parade, uit te wys deur sy/haar/hul skouer(s) aan te raak, wat (datum, tyd, plek en klagte):

....
....
....

(1) Time taken by witness to point out person(s) on parade
Tyd deur getuie geneem om persoon(e) op parade uit te wys........................

Result/Uitslag ....

....

(2) Reaction of witness during pointing out of person(s):
Reaksie van getuie tydens uitwysing van persoon(e):

....
....

(3) Comments by person pointed by witness:
Kommentaar deur persoon uitgewys deur getuie:

....
....
....

26. Suspect(s) is/are given the opportunity of changing his/her/their position(s) and asked whether he/she/they is/are satisfied.
Verdagte(s) word geleentheid gebied om van posisie(s) te verander en gevra of hy/sy/hulle tevrede is met sy/haar/hulle posisie.

His/her/their answer(s) was/were:
Sy/haar/hulle antwoord(e) was: (1)........................
(2)........................
(3) ....
(4) ....

27. Positions taken by persons on parade before the second witness appears on parade:
Posisies deur persone op parade ingeneem voordat tweede getuie op parade verskyn:

(1) .... (2)........................
(3) .... (4)........................
(5) .... (6)........................
(7) .... (8)........................
(9) .... (10)........................
(11) .... (12)........................
(13) .... (14)........................
(15) .... (16)........................
(17) .... (18)........................

28. ***Second witness/Tweede getuie:***

Name/Naam .... Language/Taal ....

was asked to point out the suspect(s), if on parade, by touching his/her/their shoulder(s), who (date, time, place and charge):
is gevra om die verdagte(s), indien op parade, uit te wys deur sy/haar/hul skouer(s) aan te raak, wat (datum, tyd, plek en klagte):

(1) Time taken by witness to point out person(s) on parade
Tyd deur getuie geneem om persoon(e) op parade uit te wys..

Result/Uitslag ....

(2) Reaction of witness during pointing out of person(s):
Reaksie van getuie tydens uitwysing van persoon(e):

(3) Comments by person pointed by witness:
Kommentaar deur persoon uitgewys deur getuie:

29. Suspect(s) is/are given the opportunity of changing his/her/their position(s) and asked whether he/she/they is/are satisfied.
Verdagte(s) word geleentheid gebied om van posisie(s) te verander en gevra of hy/sy/hulle tevrede is met sy/haar/hulle posisie.

His/her/their answer(s) was/were:
Sy/haar/hulle antwoord(e) was:
(1) ....
(2) ....
(3) ....
(4) ....

30. Positions taken by persons on parade before the second witness appears on parade:
Posisies deur persone op parade ingeneem voordat tweede getuie op parade verskyn:

(1)	(2) ....
(3)	(4) ....
(5)	(6) ....
(7)	(8) ....
(9)	(10) ....
(11)	(12) ....
(13)	(14) ....
(15)	(16) ....
(17)	(18) ....
(19)	(20) ....

31. ***Third witness/Derde getuie:***

Name/Naam .... Language/Taal ....

was asked to point out the suspect(s), if on parade, by touching his/her/their shoulder(s), who (date, time, place and charge):
is gevra om die verdagte(s), indien op parade, uit te wys deur sy/haar/hul skouer(s) aan te raak, wat (datum, tyd, plek en klagte):

(1) Time taken by witness to point out person(s) on parade
Tyd deur g`etuie geneem om persoon(e) op parade uit te wys..

Result/Uitslag ....

(2) Reaction of witness during pointing out of person(s):
Reaksie van getuie tydens uitwysing van persoon(e):

32. Remarks, if any
Opmerkings, indien enige..

33. Name of police station
Naam van polisiestasie ....
OB No.
VB No. .... /.................... /....................

34. I, No. Rank Name
Ek, No.................................... Rang.................................... Naam ....
certify that this parade was conducted by me, that the particulars which have been completed on the form by me are correct and that it is a just report of the procedures which took place.
sertifiseer dat hierdie parade deur my waargeneem is, dat die besonderhede wat op die vorm deur my ingevul, korrek en 'n presiese weergawe van die gebeure is.

....
Signature, number and rank of member in charge of parade

Appendix B
Lineup rules listed in Du Toit et al. (1987)

Eighteen guidelines for administering identification parades as listed in Du Toit et al. (1987)

Rule One: The proceedings at the parade – should at the time of the parade – be recorded (preferably on Form SAP 329) by the police official in charge of the parade.

Rule Two: The police official in charge of the parades (as referred to in Rule 1 above) should not be the investigating official, i.e., should not be the official who is charged with the investigation of the crime in respect of which the parade is held.

Rule Three: Suspects should be informed of the purpose of the parade and the allegations against them and should, further, be given an opportunity to obtain a legal representative to be present at the parade.

Rule Four: A suspect should be informed that his refusal to take part in a parade can at a possible later criminal trial be adduced as evidence against him, and, further, that the court might draw an adverse inference from such refusal or non-compliance.

Rule Five: The parade should in principle consist of at least eight to ten persons, but a greater number is desirable.

Rule Six: It is generally undesirable that there should be more than one suspect on the parade; and if a second is placed on the parade, the two suspects should be more or less similar in general appearance and the persons on the parade should be increased to at least twelve to sixteen.

Rule Seven: If the same identifying witnesses are involved in two parades, then the suspect should not be the only person appearing in both; nor should a suspect be added to a parade, already inspected by the identifying witnesses, for purposes of a second parade.

Rule Eight: The suspect and persons in the parade should be more or less of the same build, height, age and appearance and should have more less the same occupation and be more or less similarly dressed.

Rule Nine: It is extremely desirable that at least one photograph should be taken of all the persons (including the suspect) at the parade, depicting them as they appeared in the line-up and standing next to each other.

Rule Ten: The police officer in charge of the parade should inform the suspect that he may initially take up any position and change his position before any identifying witness is called.

Rule Eleven: A suspect should be asked whether he is satisfied with the parade, and, further, whether he has any requests.

Rule Twelve: It is, generally speaking, wise for the police official in charge of the parade to comply with any reasonable request made in terms of Rule 11 above, especially as regards a change of clothing.

Rule Thirteen: Identifying witnesses should be kept separately, should not be allowed to discuss the case while waiting to be called upon to attend the parade, should not be allowed to see the parade being formed or re-formed and should be kept under supervision of a police official who is neither the one in charge of the parade nor the investigating official.

Rule Fourteen: Identifying witnesses should be prevented from seeing any member of the parade before they are brought in for purposes of making an identification, and in particular should not be allowed any opportunity of seeing the suspect in circumstances indicating that he is the suspect, before or after the parade.

Rule Fifteen: A police official – who is neither the investigating official nor he official in charge of the parade nor the official charged with supervising the identifying witnesses – should escort one identifying witness at a time from the place or office where the latter is kept to the parade; and after such identifying witness's inspection of the parade, such official should escort the witness to an office or place where the witness can have no contact with witnesses who are still waiting to inspect the parade. The police official who escorts the identifying witness may not discuss the case with him.

Rule Sixteen: The supervising official referred to in Rule 13 above and the escorting official referred to in Rule 15 above, should not know who the suspect is; and the line-up should be formed and re-formed in their absence.

Rule Seventeen: The official in charge of the parade should inform each identifying witness that the person whom the witness say may or may not be on the parade and, further, that if he cannot make a positive identification, he should say so.

Rule Eighteen: The official in charge of the parade may request the identifying witness to make his identification by touching the shoulder of the suspect. And, in the event of any identification being made in this manner, it is desirable that a photo be taken of the actual act of identification.

Appendix C
Ethics Approval from the University of Cape Town

FACULTY OF HUMANITIES

PROPOSAL APPROVAL FORM

DOCTORATE (A research proposal must accompany this form)	RESEARCH MASTERS (A research proposal must accompany this form)	C/W MASTERS

SECTION A: (To be completed by candidate)
Please complete this form and return it to the Faculty Office once you have obtained the signatures of the supervisor(s) and Head of Department.

Surname	Nortje				First Name(s)	Alicia
Title	Mr.	Ms.	Mrs.	Miss	Student No	NRTALI003
Address	6A Parry Road, Claremont, Cape Town					
Telephone(Home	0216711525				Work/Cell	0836882828

Note: Your UCT Email address is the default email address for all official communication – make sure that you access it regularly.

Department	Psychology
Title of Dissertation: Investigating facial recognition for multiple-perpetrator crimes	

Qualifications held

Degree/Diploma	Major(s) &) Subjects	Month/Year awarded	University
MA	Psychological Research (coursework: Neuropsychology)	2011	UCT
Honours	Psychology	2007	UCT
BA	Psychological Counselling	2006	UNISA

Signature of candidate: ______ Date: 29 July 2013

SECTION B:

	Name	Signature	Date
Supervisor	C G Tredoux		29/07/13
Co-supervisor (if applicable)	A Vredeveldt	AVredeveldt	31/07/13
HOD	J Malcolm-Smith		31/07/13
Deputy-Dean: Research			
Ethics approval obtained where applicable	on behalf of Departmental Ethics Committee	(C. Ward)	30/7/2013

Appendix D
Ethics Approval from the South African Police Services

SOUTH AFRICAN POLICE SERVICE

SOUTH AFRICAN POLICE SERVICE

20 DEC 2016

25/7/2/1(201600143)

INFORMATION NOTE

To: The Provincial Commissioner: Western Cape

PERMISSION TO CONDUCT RESEARCH IN SOUTH AFRICAN POLICE SERVICE: THE BUTCHER, THE BAKER, THE CANDLESTICK MAKER: INVESTIGATING FACIAL RECOGNITION FOR MULTIPLE –PERPETRATOR CRIMES: PHD: UNIVERSITY OF CAPE TOWN: RESEARCHER: A NORTJE

1. The researcher, Ms Alicia Nortje, has submitted an application to conduct research within SAPS. The applicant's proposal has been perused, evaluated and recommended by the office of the Divisional Commissioner: Research.

2. The aim of the research is to determine the frequency of multiple–perpetrator crimes in South Africa, and how police interpret and implement the line–up guidelines for multi-perpetrator crimes.

3. The researcher is requesting permission to distribute surveys to SAPS members at five Police Stations in the Western Cape who have experience in conducting identification parades. The Local Criminal Record Centre will also be contacted to assist in the identifying of respondents.

4. This office has perused the application and recommends it subject to the following conditions:

   - the researcher will at his or her exclusive cost, provide all equipment of whatsoever nature used to conduct the research;
   - will conduct the research without any disruption to duties of members of the service;
   - the interviews are confined to the conducting of said interviews with members at the identified police stations;

PERMISSION TO CONDUCT RESEARCH IN SOUTH AFRICAN POLICE SERVICE: THE BUTCHER, THE BAKER, THE CANDLESTICK MAKER: INVESTIGATING FACIAL RECOGNITION FOR MULTIPLE –PERPETRATOR CRIMES: PHD: UNIVERSITY OF CAPE TOWN: RESEARCHER: A NORTJE

- prior arrangements must be made timeously with the Station Commander of such members to be interviewed to ensure that service delivery is not hampered;
- the researcher will respect the privacy of the members and will not divulge information received from a member of the Service or any person with whom the researcher conducted an interview, and that such information will at all times be treated as strictly confidential;
- If information pertains to the investigation of crime or a criminal case, the researcher must acknowledge that he or she, by publication thereof, may also be guilty of defeating or obstructing the course of justice or contempt of court;
- will pay fees or comply with further procedures in the Service, such as fees or procedures applicable to obtain access to a record of the Service;
- will allow the Service fourteen days to peruse the report in order to determine whether it complies with all conditions for the approval of the research before it is published in any manner and, if it is found not to comply with the conditions, that he or she will not publish it;
- will complete an indemnity form and agree to the undertaking and conditions prior to the commencement of his/her research, in terms of which the South African Police Service is indemnified against any injury, personal damage or any loss suffered during the research;
- the researcher may not take photographs of any office or state building as that may compromise the security of the police station, and is prohibited by law and
- Will donate an annotated copy of the research work to the Service.

____COLONEL
ACTING PROVINCIAL HEAD: ORGANISATIONAL DEVELOPMENT AND STRATEGIC MANAGEMENT
WESTERN CAPE
JAW BROWN

Date: 13/12/2016

PERMISSION TO CONDUCT RESEARCH IN SOUTH AFRICAN POLICE SERVICE: THE BUTCHER, THE BAKER, THE CANDLESTICK MAKER: INVESTIGATING FACIAL RECOGNITION FOR MULTIPLE –PERPETRATOR CRIMES: PHD: UNIVERSITY OF CAPE TOWN: RESEARCHER: A NORTJE

RECOMMENDED / ~~NOT RECOMMENDED~~

The recommendation is supported subject to compliance with the set conditions.

MAJOR GENERAL
PROVINCIAL HEAD: LEGAL SERVICES
WESTERN CAPE
FM MBEKI

Date: 2016-12-20

RECOMMENDED / NOT RECOMMENDED

MAJOR GENERAL
DEPUTY PROVINCIAL COMMISSIONER: POLICING
WESTERN CAPE
TE PATEKILE

Date: 2016-12-20

APPROVED / ~~NOT APPROVED~~

LIEUTENANT GENERAL
PROVINCIAL COMMISSIONER: WESTERN CAPE
KE JULA

Date: 2016/12/21

Appendix E
Survey administered to SAPS officers

SAPS Identification Parades Survey

We would like to invite you to participate in our survey about how police officers in the South African Police Service (SAPS) build and administer identification parades for multiple perpetrator crimes, i.e. crimes committed by a group of two or more perpetrators.

You are not forced to participate in this project, however, we would greatly appreciate your assistance if you did. You are free to withdraw your participation at any stage. If you do agree to participate, you will be asked questions about the frequency of multiple perpetrator crimes, methods that you use to build and administer multiple perpetrator/suspect identification parades, and any difficulties that you and eyewitnesses have experienced when investigating and remembering these types of crimes respectively. We are very interested in **your experience** with these parades, and we invite you to share this with us.

All the information that you provide will be anonymous and will be treated confidentially. Your responses will not be used to evaluate your job, and will not be made available to any of your superiors. A summary of the information from all the surveys will be published in the form of a journal manuscript, and we will provide a brief report to the South African Police Service, but the authors will ensure that no police officer will be identifiable from this data.

Ethical approval for this study has been granted by the Ethics Committee in the Department of Psychology at the University of Cape Town, and this data will be used towards the fulfilment of the doctoral degree of Alicia Nortje (who is listed here as one of the researchers). All questions and comments should be directed to her. Her contact details are: Alicia.nortje@gmail.com (0836882828).

If you do give consent to participate in this survey and for the researcher to collect data from you, then please provide your signature below.

I have read this consent form and give consent for my data to be used

Signature: ____

Date: ____

Instructions

In this survey you will be asked questions about your experience with identification parades, specifically those for crimes committed by multiple perpetrators (i.e., a group of two or more people). There is very little known about multiple perpetrator crimes and how police officers build and administer identification parades for such crimes. Therefore, we would like to better understand police practice surrounding these issues:

- How do police officers conduct live identification parades in South Africa?
- How are identification parades compiled when the crime has been committed by multiple perpetrators?
- Who stands in the parade with the suspect/s and how are these other people found?
- What difficulties do police officers experience as (1) the investigating officer for these crimes, and (2) the lineup administrator of these crimes?
- What suggestions do you have about how to conduct these parades?

This survey is <u>not</u> a test: We are <u>not</u> testing you to find out if you know the rules of how to conduct an identification parade. We are <u>not</u> testing you to find out if you are a good or bad officer. The information that you provide in this survey will <u>not</u> be used as a reflection of your job performance and it will not be made available your superiors.

You are an expert and a trained police officer – we do not doubt your professionalism and capabilities. The aim of this survey is to learn more about your experiences. We want to know about what you do in your job and how you solve difficult problems in investigations, and we are interested in your experiences and thoughts as an officer who is involved with identification parades.

We invite you to share your experience and knowledge with us.

Section One: Questions about experience and training

1. **Did you receive any formal training, including in-service training, about how to conduct an identification parade?** (Please tick the appropriate answer)

☐ Yes

☐ No

1.1 **If you answered yes to the previous question: What training did you receive about how to conduct an identification parade?** (If you answered 'No', then leave this question blank.)

1.2 **If you answered yes to the previous question: What was most helpful about the training** (If you answered 'No', then leave this question blank.)

1.3 **If you answered yes to the previous question: What was least helpful about the training, and were there any areas about identification parades neglected by the training?** (If you answered 'No', then leave this question blank.)

1.4 **If you answered yes to the previous question: Do you have any recommendations that you think should be added to the training on conducting identification parades?** (If you answered 'No', then leave this question blank.)

2. **For how many years have you worked for the South African Police Service?**

3. **What is your rank in the South African Police Service?**

☐ General	☐ Major
☐ Lieutenant General	☐ Captain
☐ Major General	☐ Lieutenant
☐ Brigadier	☐ Warrant Officer

☐ Colonel	☐ Sergeant
☐ Lieutenant Colonel	☐ Constable

4. **Roughly how many years of experience do you have with <u>conducting / building / administering</u> identification parades?**

5. **Roughly how many identification parades have you formed/administered across your career?**

☐ 5 or fewer (Please provide an estimate: ____________________________________)
☐ Between 5 and 10 (Please provide an estimate: ____________________________________)
☐ Between 10 and 25 (Please provide an estimate: ____________________________________)
☐ Between 25 and 50 (Please provide an estimate: ____________________________________)
☐ More than 50 (Please provide an estimate: ____________________________________)

6. **Roughly how many identification parades have you formed/administered in the last 12 months (if none or zero, then please write that)?**

7. **Have you ever testified in court about an identification parade that you formed and administered or used in your investigation?**

☐ Yes

☐ No

7.1 If you answered 'yes' to question 7, then please roughly estimate how many times you been called to court to testify on an identification parade that you formed or used in your investigation across your career (if never, then please state write that)?

Section Two: Investigating crimes

(Please answer all of these questions about crimes that you have investigated.)

8. **What types of crime do you investigate the most often/specialise in? You can include more than one type, especially if you have specialised in many different types of crime in your career.**

9. **Of the crimes that you investigated in the last 12 months, what percentage was committed by multiple (2 or more) perpetrators/criminals?** (Please circle the appropriate answer.)

0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%

10. **Of the crimes that you investigate and/or specialise in, how many perpetrators are usually involved/or commit these types of crimes? If you specialise in more than one type of crime, please list answers for all of them.**

11. **Of the crimes that you investigate and/or specialise in, what is the greatest number of perpetrators that have committed a single crime? If you specialise in more than one type of crime, please list answers for all of them.**

12. **In your experience, which types of crime, including those that you specialise in, are more likely to be committed by multiple perpetrators:**

Types of crime most likely to be committed by multiple perpetrators:

12.1 In your experience, are there any other crimes that are typically committed by multiple perpetrators that I have not listed?

Section Three: Building parades

(Please answer the following questions about the identification parades that you have constructed and administered.)

13. Have you ever administered or built an identification parade that contained more than one suspect?

☐ Yes

☐ No

14. In your experience, which have you constructed most frequently: parades that contain only <u>one</u> suspect **or** parades that contain <u>two or more</u> suspects (please tick which occurs more frequently)?

☐ Parades that contain only one suspect

☐ Parades that contain two or more suspects

15. In your experience as the <u>investigating</u> officer, when, if ever, would you decide to include two or more suspects in one parade?

16. Imagine a situation where a single crime is committed by two or more people. If you are going to build a parade for this crime, and you have two or more suspects who committed this crime together, which method would <u>you prefer</u> to use:

☐ Method One: Build a single parade that contains all the suspects and other innocent people. Therefore in this example, you would have a single parade that contains two or more suspects and other innocent people.

☐ Method Two: Build as many parades as there are suspects, so that each parade contains only one suspect and other innocent people. Therefore you would build two parades, and the first suspect will appear in the one parade and the second suspect will appear in the second parade.

☐ Method Three: Is there a different method that you would use? Please describe it here:

16.1 Please provide reasons for why you prefer the Method that you chose in question 15:

17. Of all the parades held for crimes that you have investigated/administered, what is the greatest number of <u>suspects</u> (for one crime) that you have placed together in the same, single parade?

18. **In your experience, what is a realistic, obtainable number of innocent people (i.e. the other line-up members) to include in a parade if:**
    (Please write the appropriate number)
    There is one suspect in the parade: __________________
    There are two or more suspects in the same parade: ____________________

19. **What is the largest parade (suspects and other line-up members) that you have ever built/administered?**

20. **In your experience, how are the other (i.e. innocent people) people who appear in the parade alongside the suspect found?**

21. **Before the parade is formed, does the suspect have any input in finding/sourcing the people who will appear in the parade with them?** (If yes, please describe what input they have)

22. **In your experience, do you more often arrange parades for crimes with multiple *eyewitnesses* (two or more) or single *eyewitnesses*?**
    - ☐ Single eyewitness
    - ☐ Two or more eyewitnesses

23. **In your experience, how much time on average would have passed between when the crime took place and when an identification parade is held?**

24. **Have you used any of the following parades? (**Please tick all the options that you have used.**)**

☐ A live parade
☐ A photograph parade
☐ A video parade
☐ Other type of parade (please describe it here)

25. If you could choose from the types of parades above, which would you prefer and why?

Which parade do you prefer: Why do you prefer this parade?

26. When an eyewitness makes identification from the line-up, are they required to state a reason for why/how they made their identification?

☐ Yes

☐ No

27. When an eyewitness makes an identification from the line-up, are they required to state any additional information about the person (e.g. "I recognise him because he had the gun").

☐ Yes

☐ No

28. In your experience as the officer administering the parade, are eyewitnesses able to name and describe the roles/actions (e.g. "He had the gun") of *all* the people whom they identify from the parade?

☐ Yes

☐ No

29. In your experience as the officer administering the parade, do eyewitnesses ever seem confused about the roles and actions of the people whom they identify from the identification parade (that is, they are not certain *who did what* but they are certain that they were involved)?

☐ Yes

☐ No

30. In your experience as the investigating officer, has it ever happened that the information that the eyewitness provided when making an identification (like in Question 26, 27, 28 and 29) *differed* from the information that they provided in their statement?

- ☐ Yes
- ☐ No

30.1. If you answered yes to the previous question (Question 30):
How did the information that the eyewitness provided when making an identification differ from the information in their statement ?(If you answered 'No', then leave this question blank.)

31. **Based on the different questions that I have asked here, is there any other important information that you think we should know about:**

Building parades with multiple suspects:

Administering parades to eyewitnesses:

Difficulties experienced by eyewitnesses of multiple perpetrator crimes:

Difficulties experienced by police officers who administer/form identification parades:

Anything else:

Appendix F
Old and New Attributes

The List of Old and New Attributes used in Face Recognition Experiment 1 (Chapter 4).

Old Attributes (used at encoding)	#	Matched Attributes	#
He's a skydiving instructor.	1	He's a paragliding instructor.	1
He hates raisins.	2	He hates prunes	2
He broke his arm rowing.	3	He broke his arm surfing.	3
He loves playing Playstation.	4	He loves playing Xbox.	4
He collects coins.	5	He collects stamps.	5
He eats mayonnaise with everything.	6	He eats tomato sauce with everything.	6
He plays the trumpet.	7	He plays the saxophone.	7
He sings in a church choir.	8	He sings in a youth group.	8
He's one of an identical twin.	9	He's one of a triplet.	9
He went to Namibia for the holidays.	10	He went to the Kalahari for the holidays.	10
He's allergic to peanuts.	11	He's allergic to almonds.	11
He cycles twice a day.	12	He gyms twice a day.	12
He speaks Italian.	13	He speaks Spanish.	13
He loves eating naartjies.	14	He loves eats oranges.	14
He only smokes cigars.	15	He only smokes pipe.	15
He draws cartoons.	16	He draws portraits.	16
He is a stand-up comedian.	17	He is a stage actor.	17
He owns chickens.	18	He owns ducks.	18
He plays online poker.	19	He plays online blackjack.	19
He's written a novel.	20	He's written a film screenplay.	20
He makes his own beer.	21	He makes his own wine.	21
He bakes his own biscuits.	22	He bakes his own rusks.	22
He volunteers at the local old age home.	23	He volunteers at the local night shelter.	23
He speaks with a lisp.	24	He speaks with a stutter.	24
He wanted to be an astronaut.	25	He wanted to be a pilot.	25
He's terrified of spiders.	26	He's terrified of snakes.	26
He avoids food that contains milk.	27	He avoids food that contains wheat.	27
He has three tattoos.	28	He has three piercings.	28
He loves to play squash.	29	He loves to play tennis.	29
He drives a Tazz	30	He drives a Golf	30

Appendix G
Orders and Pairing of Faces and Attributes

Two Predefined Orders (Order A, and Order B) of Attributes and Faces in Face Recognition Experiment 1.

Order A		Order B	
Attribute	Face	Attribute	Face
He's a skydiving instructor.	24	He is a stand-up comedian.	9
He hates raisins.	22	He loves eating naartjies.	15
He broke his arm rowing.	23	He only smokes cigars.	7
He loves playing Playstation.	20	He speaks Italian.	11
He collects coins.	8	He draws cartoons.	4
He eats mayonnaise with everything.	13	He volunteers at the local old age home.	18
He plays the trumpet.	29	He's allergic to peanuts.	19
He sings in a church choir.	6	He avoids food that contains milk.	14
He's one of an identical twin.	30	He eats mayonnaise with everything.	13
He went to Namibia for the holidays.	16	He's terrified of spiders.	25
He's allergic to peanuts.	19	He speaks with a lisp.	12
He cycles twice a day.	10	He plays the trumpet.	29
He speaks Italian.	11	He drives a Tazz	28
He loves eating naartjies.	15	He's written a novel.	17
He only smokes cigars.	7	He broke his arm rowing.	23
He draws cartoons.	4	He's one of an identical twin.	30
He is a stand-up comedian.	9	He loves to play squash.	26
He owns chickens.	5	He bakes his own biscuits	3
He plays online poker.	21	He's a skydiving instructor.	24
He's written a novel.	17	He owns chickens.	5
He makes his own beer.	2	He loves playing Playstation.	20
He bakes his own biscuits.	3	He makes his own beer.	2
He volunteers at the local old age home.	18	He plays online poker.	21
He speaks with a lisp.	12	He went to Namibia for the holidays.	16
He wanted to be an astronaut.	1	He hates raisins.	22
He's terrified of spiders.	25	He sings in a church choir.	6
He avoids food that contains milk.	14	He has three tattoos.	27
He has three tattoos.	27	He cycles twice a day.	10
He loves to play squash.	26	He collects coins.	8
He drives a Tazz	28	He wanted to be an astronaut.	1

Appendix H
Number of Trials

Number of trials per set size group in Face Recognition Experiment 1 (Chapter 5)

Set Size Group	Number of face-attribute pairs studied per trial	Number of Trials (Encoding – Recognition)	Total number of pairs of faces and attributes studied
1	1	30	30
2	2	15	30
3	3	10	30
5	5	6	30
10	10	3	30
15	15	2	30
30	30	1	30

Appendix I
List of Old and New Attributes

A List of Old and New Attributes used in Face Recognition Experiment 2 (Chapter 5)

Item	Old Attributes (shown at encoding)	New, Matched Attributes
1	He sings in a church choir.	He sings in a youth group.
2	He bakes his own biscuits.	He bakes his own rusks.
3	He draws cartoons.	He draws portraits.
4	He’s allergic to peanuts.	He’s allergic to almonds.
5	He speaks with a lisp.	He speaks with a stutter.
6	He loves to play squash.	He loves to play tennis.
7	He collects coins.	He collects stamps.
8	He avoids food that contains milk.	He avoids food that contains wheat.
9	He’s written a novel.	He’s written a film screen play.
10	He owns chickens.	He owns ducks.
11	He only smokes cigars.	He only smokes pipe.
12	He broke his arm rowing.	He broke his arm surfing.
13	He’s terrified of spiders.	He’s terrified of snakes.
14	He’s one of an identical twin.	He’s one of a triplet.
15	He eats mayonnaise with everything.	He eats tomato sauce with everything.
16	He speaks Italian.	He speaks Spanish.
17	He went to Namibia for the holidays.	He went to the Kalahari for the holidays.
18	He plays the trumpet.	He plays the saxophone.
19	He cycles twice a day.	He gyms twice a day.
20	He has three tattoos.	He has three piercings.
21	He wanted to be an astronaut.	He wanted to be a pilot.
22	He makes his own beer.	He makes his own wine.
23	He loves playing Playstation.	He loves playing Xbox.
24	He hates raisins.	He hates prunes.

Appendix J
Orders of Faces and Facts

Two Orders (Order A, Order B) of Faces and Facts for Face Recognition Experiment 2

Order A		Order B	
Faces	Facts	Faces	Facts
01E19	01F06	01E08	01F06
02E05	02F03	02E12	02F03
03E21	03F04	03E32	03F04
04E18	04F19	04E18	04F19
05E13	05F12	05E24	05F12
06E17	06F26	06E20	06F26
07E32	07F08	07E05	07F08
08E20	08F14	08E25	08F14
09E24	09F17	09E09	09F17
10E22	10F05	10E04	10F05
11E09	11F07	11E26	11F07
12E25	12F23	12E19	12F23
13E20	13F25	13E27	13F25
14E11	14F30	14E31	14F30
15E08	15F13	15E11	15F13
16E31	16F11	16E17	16F11
17E28	17F16	17E01	17F16
18E29	18F29	18E28	18F29
19E16	19F10	19E13	19F10
20E04	20F27	20E22	20F27
21E12	21F01	21E30	21F01
22E27	22F02	22E29	22F02
23E01	23F20	23E21	23F20
24E26	24F22	24E16	24F22

Appendix K
Remember and Know Instructions

Remember and Know Instructions given to participants in Face Recognition Experiment 2. These instructions were revised to suit the stimuli (faces, facts, pairs) used in the experiment.

Slide 1:

If you recognise the face, I want you to also consider whether you 'Remember' the face or 'Know' the face. if you do not recognise the face, then the face is 'New'.

Often when remembering a previous event, we also consciously recollect and become aware of aspects of the previous experience. Conscious recollection would include the ability to become consciously aware again of some aspect of what had happened and what was experienced at the time the face was presented - such as something to do with the physical appearance of this face, the way it was presented, something one was thinking of or did during the face's presentation, or something else that you noticed in the laboratory at that time.

We will refer to this as 'Remembering'

On the other hand, sometimes we simply know that something has occurred before, but without being able consciously to recollect anything about its occurrence of what we experienced at the time. For example, recognising that you had studied the face, but being unable to consciously recollect anything about the actual occurrence of the face, or what happened and what was experienced at the time it was presented.

We will refer to this as 'Knowing'.

Please press 'SPACEBAR' to continue.

Slide 2:

To further explain the difference between Knowing and Remembering, consider this:

Knowing does not mean that you have worse memory for the event - for example, if I ask you for your name, you can recall it correctly without consciously recalling any events or experiences.

But if I ask you about the last film you watched, your memory would probably be accompanied by consciously recalling some aspects of that film and the experience.

So to summarise: If you have a Remember-response to the face, then you have to be able to describe something specific that you remember about when you studied that specific face. If you are not able to recall anything specific about when you studied that specific face, but you do recognise it as one that you studied previously, then this is a Know-response. If you do not recognise the face as one that you studied earlier and you think it was new, then respond New.

Press 'SPACEBAR' to continue to the next set of instructions.

Appendix L
Sample Image from Eyewitness Encoding Video

Still image taken from single-perpetrator video (Eyewitness Experiment, Chapter 5).

Appendix M
Eyewitness Questionnaire

Questions included in the statement for the Eyewitness Experiment (Chapter 5).

Statement questions for single-perpetrator condition:

1. How many people were involved in the crime?
2. Please describe the physical appearance of the person who committed the crime. Consider that your description may be used to find and identify this person at a later stage.
3. What did the perpetrator do in the video of the crime?
4. Please describe what the perpetrator did in the form of a time-line from the beginning until the end of the video.
5. Was anything stolen from the laboratory?

Statement questions for multiple-perpetrator condition:

1. How many people were involved in the crime?
2. Please describe the physical appearance of each person who committed the crime. Consider that your description may be used to find and identify these people at a later stage.
3. What did each perpetrator do in the video of the crime?
4. Please describe what each perpetrator did in the form of a time-line from the beginning until the end of the video.
5. Was anything stolen from the laboratory?
6. Did one of the perpetrators appear to be in charge?

Appendix N
Lineup Instructions

Lineup instructions for the Eyewitness Experiment (Chapter 5).

Lineup instructions for single-perpetrator condition:

You will be presented with a lineup consisting of six men.
Your task is to identify which of these men was the perpetrator in the film that you watched, if he is present in the lineup.
Please make your choice by pressing the number below his head; if you do not know or do not recognise him, please press 0.
Please press SPACEBAR to continue.

Lineup instructions for multiple-perpetrator condition:

You will be presented with several lineups consisting of six men.
Your task is to identify which of these men was one of the perpetrators in the film that you watched, if he is present in the lineup.
Please make your choice by pressing the number below his head. If you do not know or do not recognise him, please press 0.
You will not be able to revisit previous lineups to change your decisions.
Please press SPACEBAR to continue.

Appendix O
Difficulties with Calculating Within-Subject Confidence Intervals for Eyewitness Experiment

I was not able to calculate within-subject confidence intervals for this data. To calculate the proportion of hits, and the proportion of accurate pairs, I summed the number of correct responses within each group, and divided this by the number of TP, and Hits, respectively; thus, the base unit of measurement was the number of accurate responses. I investigated calculating within-subject confidence intervals using the methods described by Loftus and Masson (1994) and Morey (2008) – which is a revision of the method proposed by Cousineau (2005). An advantage to using the Cousineau-Morey method, is that it accounts for the number of trials that participants complete. For Cousineau-Morey confidence intervals, the standard CI formula is multiplied by the square-root of (c/c-1), where c is the number of within-subject trials. For Set Sizes 1, 2, and half of participants in Set Size 3, participants only complete one trial, thus this correction suggested by Morey cannot be calculated (the correction would have a denominator of zero). I decided to return to Loftus and Masson (1994) method. For Loftus and Masson (1994), participant-means are normalising by subtracting from the grand mean; the grand mean, however, is not representative as the number of trials differed between subjects. Thus, the grand mean cannot be calculated from the number of hits, and the number of accurate pairs, as the base units have a rank and differ between groups. To standardise the measurements between groups, it would be better to use the proportion. The base unit can be transformed into a proportion for each participant, and a grand mean can be calculated from the proportion. However, if one does this, then the group means are different: For example, calculating the mean for Set Size 3 using participant-wise proportions versus the sum of accurate pairs/sum of Hits is .69 and .76 respectively, leading to different CIs. The difference arises from the proportion-value for each unit within each set size, for example, one accurate pair out of two Hits in Set Size 3 has a proportional value of .50, but one accurate pair out of a total of 29 Hits has a proportional vale of .03. Furthermore, each accurate unit has a different proportional value between group; consequently, within-group variability for proportions varies greatly between groups. (e.g., participants in Set Size 2, can one TP parade; Set Size 3 have one or two TP parades; Set Size 5 have two or three TP parades; Set Size 10 has five TP parades). I investigated the method outlined in Jarmasz and Hollands (2009), and considered using the MSE for the mixed effects logistic regression. The MSE, however, is based on the logit, and not based on the original values from the proportions that were calculated. My ultimate decision was that I would report the mixed effects logistic regression with the confidence intervals, and that this would aid the reader, and rather report the figure (where the proportion is calculated from the sum of the base units, rather than within-participant proportions) so that this figure could mirror the pattern of results of the mixed effects logistic regression.

Appendix P
Parameters for Revised IAC model

All three models had the following global parameters, which were set at the following values:

Global Parameter	Values
Maximum activation	1.0
Minimum activation	-0.2
Resting activation	-0.1
Decay rate	0.09
Strength of external input	0.4
Alpha (strength of excitatory input)	0.1
Gamma (strength of inhibitory input)	0.1

Excitatory connections had a weight of 1.0.
Inhibitory connections had a weight of -0.1.
The number of cycles was set to 100.