A Re-examination of Eyewitness Memory Phenomena Using Receiver Operating Characteristic Analysis and Confidence-Accuracy Characteristic Analysis

Travis Morgan Seale-Carlisle

Thesis submitted in fulfilment of the degree of

Doctor of Philosophy

Department of Psychology

Royal Holloway, University of London

Abstract

Eyewitness identification research is focused on investigating the factors that affect the eyewitnesses' ability to accurately identify the perpetrator from a lineup. A lineup consists of the police suspect and several other individuals who resemble the perpetrator, but are known to be innocent (called fillers). Several decades of research continue to have a growing impact on criminal justice systems throughout the world, most notably in the US, by informing public policy and informing the court (i.e. judges and jurors) through expert testimony. Efforts to shape public policy have been directed at improving fundamental aspects of the identification process with the goal to implement procedures that maximize discriminability – the ability to distinguish innocent from guilty suspects. Yet, poor measures of discriminability have resulted in many US jurisdictions implementing substandard procedures that may actually *reduce* discriminability. In court, factors that reduce discriminability are believed by many experts to reduce the reliability of an eyewitness identification – the accuracy of a suspect identification admitted as evidence in court. However, discriminability and reliability are two separate measures of eyewitness identification "accuracy." That is, an eyewitness may have poor discriminability, but may nonetheless make a reliable identification from a lineup. To critically assess this issue, I have re-examined several eyewitness memory phenomena including the sequential superiority effect (Chapter 3), the verbal overshadowing effect (Chapter 4), and the weapon focus effect (Chapter 5) using two analytic techniques recently introduced to the eyewitness identification field that measure discriminability and the reliability of a suspect identification: receiver operating characteristic analysis and confidence-accuracy characteristic analysis, respectively. Together, this research highlights the importance of distinguishing between discriminability and reliability. Appreciating this distinction can help inform policymakers of procedures that boost discriminability and can help inform the court of the reliability of a suspect identification.

Declaration of Authorship

I, Travis M. Seale-Carlisle, hereby declare that this work was carried out in accordance with the Regulations of the University of London. I declare that this submission is my own work, and does not represent the work of others, published or unpublished, except where duly acknowledged in the text. No part of this thesis has been submitted for a higher degree at another university or institution.

Signed _____

Dated _____

Acknowledgements

I would first like to thank my supervisor, Dr. Laura Mickes. You have helped me grow as a researcher and as a person over the last eight years. I cherish our relationship. I know that without your guidance, I would not have made it this far. I sincerely thank you for supervising me.

I would also like to thank my family for all of their support throughout the completion of this thesis. To my Dad, thank you for your overwhelming support. I know that you will agree that my future is “so bright” that I need sunglasses (true dat, double true). Love you Dad. To my Mom, you have continued to support me through thick and thin. I love you. You are the best mom in the world. I would especially like to thank my brother, Justin Seale-Carlisle, for always being there for me and continuing to be someone who I look up to and admire. You deserve everything in life. I can’t even imagine where I would be without you. You’re an amazing brother. Love you bruv!

I would also like to thank Gurpreet. You are my partner through and through. I love you, even though you are not very good at FIFA and you get scared easily. I look forward to our future together. I’m proud of you and greatly admire your strength and charm. You can always make me laugh and smile.

Lastly, I would like to thank all of my friends in the Psychology Department for being kind, resourceful, and thoughtful throughout my time here. Science is hard, but each of you helped to make it a bit easier.

Table of Contents

List of Tables 10
List of Figures 11
Chapter 1: Introduction 13
The Basic Identification Procedure 13
Consequences of a False ID 14
Consequences of a Miss 14
Standard Eyewitness Identification Experiment 14
Assisting the Criminal Justice System: Experts in Court 15
Should Eyewitness Experts Testify? 16
Eyewitness Expert in Court 16
Measuring the Reliability of a Suspect Identification 17
Assisting the Criminal Justice System: Consultants to Policymakers 24
Review of System Variable Research 25
Misuse of the Diagnosticity Ratio 26
Receiver Operating Characteristic Analysis 29
Signal-Detection Theory 33
Diagnostic-Feature-Detection Hypothesis 39
Preview of Upcoming Chapters 39
Chapter 2: General Methods 40
General Procedure 40
Calculating Correct ID, False ID, and Filler ID rates. 41
Correct ID Rate 41
False ID Rate 41
Filler ID Rate 42
Receiver Operating Characteristic Analysis 42
Statistically Comparing *pAUC* Values 43

Estimating Lineup Discriminability 43
Testing for Significance 47
Comparing *d'* and *pAUC* 47
Confidence-Accuracy Characteristic Analysis 47
Suspect ID Accuracy and Prior Probability 48
Computing CAC Standard Errors 49
Chapter 3: The US Lineup Outperform UK Lineup 50
UK Lineup 50
US Lineup 51
US Lineup vs. UK Lineup Predictions 52
Moving vs. Static Images 53
Multiple Laps vs. One Viewing 54
Nine vs. Six Lineup Members 56
US vs. UK Predictions Revisited 59
Confidence-accuracy Characteristic Analysis 59
Experiment 1 and 2 60
Methods 60
Results 61
General Discussion 67
Chapter 4: The Effect of Descriptions on Discriminability and Reliability 69
Replicating the Verbal Overshadowing Effect 69
Theoretical Implications of Verbal Overshadowing 70
Content Accounts of Verbal Overshadowing 70
Recoding Interference 70
Retrieval-based Interference 72
Content Account Limitations 73
Transfer-Inappropriate Processing Account 74
Transfer-appropriate Processing Framework 75

Retrieval-induced Forgetting Framework 76
Evidence for Transfer-inappropriate Processing 77
Criterion Shift Account 79
Criterion Shift Account Limitations 80
Experiment 3 81
Reliability of an Identification 82
Methods 82
Results 84
Discussion 87
Experiment 4 90
Methods 91
Results 92
Discussion 95
Experiment 5 98
Methods 99
Results 100
General Discussion 106
Chapter 5: The Effect of a Weapon on Discriminability and Reliability 108
Weapon Focus in the Laboratory 108
Arousal/Stress Hypothesis 108
Unusual Item Hypothesis 111
Weapon Focus for Actual Crimes 114
Expert Opinion 115
Gaps in Research 116
Perceptual Analysis 116
Is Weapon Focus Meaningful to Judges and Jurors? 118
Experiment 6 119
Methods 119

Results 121
Discussion 126
Experiment 7 128
Methods 128
Results 129
General Discussion 133
Chapter 6: Confidence is the Best Available Marker of Suspect ID Accuracy 135
Confidence in the Identification Decision 135
Visual Behaviour 137
Visual Behaviour during the Study Phase 137
Visual Behaviour during the Recognition Phase 138
Visual Behaviour in Eyewitness Studies 138
Pupil Dilations 140
Pupil Dilations and Recognition Memory 140
Pupil Dilations and Decision-Making 143
The Present Experiment 144
Methods 145
Results 147
General Discussion 151
Chapter 7: General Discussion 153
Chapter 3 154
Experiment 1 and 2 154
Chapter 4 155
Experiment 3 and 4 155
Experiment 5 157
Chapter 5 159
Experiment 6 and 7 159

Chapter 6 161
Experiment 8 161
Limitations and Future Research 162
Concluding Statement 164
References 165
Appendix 186
Chapter 3 186
Chapter 4 189
Chapter 5 190

List of Tables

Chapter 1: Introduction

Table 1: The six decision outcomes of a lineup……13

Table 2: The four decision outcomes of a recognition memory experiment…….....34

Chapter 3: The US Lineup Outperforms the UK Lineup

Table 1: The difference between the UK and US Lineups……....52

Table 2: Nominal and functional lineup size…….58

Table 3: Response frequencies for Experiment 1 and 2……62

Table 4: Identification rates for Experiment 1 and 2…….64

Chapter 4: The Effect of Descriptions on Discriminability and Reliability

Table 1: Response frequencies for Experiment 3……..84

Table 2: Identification rates for Experiment 3 ……..85

Table 3: Response frequencies for Experiment 4……..92

Table 4: Identification rates for Experiment 4 ……..93

Table 5: Response frequencies for Experiment 5…….101

Table 6: Identification rates for Experiment 5 ……103

Chapter 5: The Effect of a Weapon on Discriminability and Reliability

Table 1: Weapon focus trend for each crime video…….. 121

Table 2: Response frequencies for Experiment 6……....123

Table 3: Identification rates for Experiment 6 ……124

Table 4: Weapon focus trend for each crime video…….. 129

Table 5: Response frequencies for Experiment 7……....130

Table 6: Identification rates for Experiment 7 ……131

Chapter 6: Confidence is the Best Available Marker of Suspect ID Accuracy

Table 1: Response frequencies for Experiment 8……....147

List of Figures

Chapter 1: Introduction

Figure 1: Measuring reliability……..……23
Figure 2: Comparing outcomes for simultaneous and sequential lineups…….....31
Figure 3: Simultaneous lineup yields greater discriminability than sequential lineup….33
Figure 4: Equal-variance signal-detection model for recognition memory..…….....35
Figure 5: Shift in response bias…….……36
Figure 6: Shift in discriminability…….....37
Figure 7: Equal-variance signal-detection model for lineups……38

Chapter 2: Methods

Figure 1: Equal-variance signal-detection model for lineups……....44
Figure 2: Visual depiction of d' ……..……..……46

Chapter 3: The US Lineup Outperforms the UK Lineup

Figure 1: Receiver operating characteristic analysis for Experiments 1 and 2...…….....65
Figure 2: Confidence-accuracy characteristic analysis for Experiments 1 and 2……...67

Chapter 4: The Effect of Descriptions on Discriminability and Reliability

Figure 1: Alogna et al. (2014) RRR1 task procedure…….……..…..84
Figure 2: Receiver operating characteristic analysis for Experiment 3……....……....86
Figure 3: Confidence-accuracy characteristic analysis for Experiment 3…….…….87
Figure 4: Alogna et al. (2014) RRR2 task procedure…….……..…..91
Figure 5: Receiver operating characteristic analysis for Experiment 4……....……....94
Figure 6: Confidence-accuracy characteristic analysis for Experiment 4…….…….95
Figure 7: Receiver operating characteristic analysis for Experiment 5……....……...105
Figure 8: Confidence-accuracy characteristic analysis for Experiment 5…….…...106

Chapter 5: The Effect of a Weapon on Discriminability and Reliability

Figure 1: Receiver operating characteristic analysis for Experiment 6……....……...125
Figure 2: Confidence-accuracy characteristic analysis for Experiment 6…….…...126
Figure 3: Receiver operating characteristic analysis for Experiment 7……....……...132

Figure 4: Confidence-accuracy characteristic analysis for Experiment 7……..…...133

Chapter 6: Confidence is the Best Available Marker of Suspect ID Accuracy

Figure 1: Confidence-accuracy characteristic analysis for Experiment 8……..…...148

Figure 2: Pupil dilation ratios for item type and decision outcomes…….……..149

Figure 3: Pupil dilation ratios across confidence ratings…….……151

Chapter 1

The Basic Identification Procedure

In criminal investigations police commonly administer a lineup procedure to an eyewitness of a crime in order to identify the perpetrator or exclude an innocent person who is suspected by police. A lineup consists of a police suspect who is either guilty or innocent and several other individuals who match the general description of the perpetrator but are known to be innocent (called fillers or foils). The eyewitness may identify a lineup member or identify no one, resulting in six possible decision outcomes. Each of these six possible outcomes is presented in Table 1. There are two possible correct responses: 1) a correct identification (correct ID), and 2) a correct rejection. Successfully identifying the perpetrator from a perpetrator-present lineup (also known as a target-present lineup) is known as a correct ID. Rejecting the lineup when the perpetrator is not in the lineup (i.e., from a perpetrator-absent lineup, also known as a target-absent lineup) is known as a correct rejection. There are three possible incorrect responses: 1) a false ID, 2) a miss, and 3) a filler ID. A false ID occurs when the eyewitness identifies the innocent suspect from a perpetrator-absent lineup. A miss occurs when the eyewitness fails to identify the perpetrator from a perpetrator-present lineup (incorrectly rejecting the lineup instead). The eyewitness may also identify a filler from a perpetrator-present or perpetrator-absent lineup, but because fillers are known to be innocent, they are not at risk of wrongful conviction. However, false IDs place the innocent suspect at risk of wrongful investigation and possible conviction and a miss may result in a guilty suspect going free. Thus, false IDs and misses are the two errors that are of great practical importance because both of these errors may result in harm.

Table 1

The six decision outcomes of a lineup.

	Suspect ID	No ID	Filler ID
Perpetrator-Present	Correct ID	Miss	Filler ID
Perpetrator-Absent	False ID	Correct Rejection	Filler ID

Consequences of a False ID

Perhaps the most well-known and well-documented example of a false ID is the case of Jennifer Thompson. In July 1984, Thompson was raped in her apartment. Later that night, police apprehended Ronald Cotton, a young man that fit Thompson's description of the perpetrator. Thompson first identified Cotton out of a photo lineup. Later, Thompson stated that she was fairly sure Cotton was the perpetrator but admitted that the task was very difficult as all of the lineup members matched her description. During the live lineup several days later, Thompson vacillated between choosing Cotton and one other individual, and over time, she identified Cotton with even more certainty. At the trial, Thompson emphatically declared Cotton as her rapist. "I was absolutely, positively, without-a-doubt certain he was the man who raped me when I got on that witness stand and testified against him," Thompson recalls, "and nobody was going to tell me any different" (Hansen, 2001). Thompson's absolute certainty in court persuaded jurors to convict Cotton, who was subsequently sentenced to life plus 50 years in prison. After serving nearly 11 years, DNA evidence established Cotton's innocence (Thompson-Cannino, Cotton, & Torneo, 2009).

Consequences of a Miss

A photograph of the notorious serial killer, Ted Bundy, was presented to an eyewitness who had seen Bundy kidnap two women in broad daylight (Kendall, 1981). After browsing through a series of photographs that included a photo of Bundy and several other fillers, the eyewitness failed to identify Bundy as the perpetrator. Within the following years, Bundy had killed several more women before his final arrest.

Standard Eyewitness Identification Experiment

False IDs place innocent suspects at risk of wrongful investigation and possible conviction. A miss may remove suspicion from a perpetrator who is a threat to public safety. Because false IDs and misses may result in harm, it is critical that both types of errors are reduced (Ebbesen & Flowe, 2002; Mickes, Flowe, & Wixted, 2012). Of course, in a police-constructed lineup it is unknown whether or not the police suspect is guilty and

so, it is difficult to determine whether a false ID or a miss has occurred. However, in the laboratory, we can manipulate guilt. That is, experimenters assign guilt or innocence to a suspect. In the typical eyewitness memory experiment, participants take on the role of an eyewitness by watching a video of a mock crime and attempting to identify the perpetrator from a perpetrator-present or perpetrator-absent lineup. The number of correctly identified perpetrators divided by the number of perpetrator-present lineups is known as the correct ID rate and the number of innocent suspects falsely identified divided by the number of perpetrator-absent lineups is known as the false ID rate. By simulating the eyewitness identification procedure in a controlled laboratory (where guilt and innocence can be definitively established), researchers are able to elucidate the factors that reduce the rates of false IDs and misses. This research can then help inform the criminal justice system.

Eyewitness identification researchers have been able to inform the criminal justice system largely through two streams: as eyewitness experts in the courts of law and as consultants to policymakers. As eyewitness experts in the courts of law, triers of fact (i.e. judges and jurors) can become informed of the multitude of factors that affect the reliability of suspect IDs. By consulting with policymakers, police procedures that reduce the rates of false IDs and misses can be implemented.

Assisting the Criminal Justice System: Experts in Court

Eyewitness researchers have, for quite some time, sought to assist and improve upon the criminal justice system. In the early 1900s, Munsterberg (1908) was convinced that eyewitness testimony often contained errors and that psychological science was best equipped to inform the court of these errors. Attorneys and legal scholars refuted this suggestion by criticizing the lack of practical solutions psychological science could provide (e.g. Moore, 1907, Wigmore, 1909). Several decades later however, during the mid-1970s, psychological researchers began to demonstrate the fallibility of eyewitness memory (e.g. Loftus & Palmer, 1974) and soon after, defense attorneys began to call on eyewitness researchers in court to challenge or, on occasion, strengthen key testimony (Egeth, 1993).

Should Eyewitness Experts Testify?

Some eyewitness identification researchers questioned whether the results from the laboratory were generalizable to the "real world" (Yuille & Cutshall, 1986), whether there was adequate scientific foundation for such testimony (Konecni & Ebbesen, 1986), or if averaged data across a group of participants taking on the role of eyewitnesses in a laboratory could be applied to an actual eyewitness identification from a single case (Memon, Mastroberardino, & Fraser, 2008; Wells et al., 2000). There was also concern for the potential of expert testimony to invade the province of the jury (i.e. the power of the jury to decide on the facts). Critics preferred traditional safeguards such as cross-examination of the eyewitness or the use of judicial instructions as an aid to jury decision-making instead (Penrod & Cutler, 1989). Despite these concerns, many psychological researchers (e.g. Kassin, Ellseworth, & Smith, 1989; Leippe, 1995; Loftus, 1983), lawyers (Frazzini, 1981; Stein, 1981; Woocher, 1977), and judges (Bazelon, 1980; Weinstein, 1981) defended its use in court as a tool to improve jury decision-making. Since the mid-1980s, appellate courts have been more receptive to admitting eyewitness expert testimony (e.g. People v. McDonald, 1984; State v. Chapple, 1983; State v. Moon, 1986) and, consequently, many courts throughout the United States have become increasingly open to such testimony (Wells et al., 2000).

Eyewitness Expert in Court

Prior to criminal trials, eyewitness experts have helped attorneys examine eyewitness evidence and have helped attorneys present the eyewitness evidence to the trier of fact (e.g. Loftus & Ketchum, 1992). During criminal trials, eyewitness experts have provided testimony in court to help the trier of fact recognize when an error in testimony has likely occurred (Constandi, 2013; Costanzo, Krauss, & Pezdek, 2007; Kassin et al., 1989; 2001; Leippe, 1995; Loftus, 1983). When a suspect ID has been admitted as evidence in court, eyewitness experts have sought to provide a context or framework for evaluating the reliability of suspect IDs and have refrained from commenting on the reliability of the particular suspect ID of the case (doing so would invade the province of the jury) (Costanzo et al., 2007; Monohan & Walker, 1988).

Measuring the Reliability of a Suspect Identification

The reliability of a suspect ID admitted as evidence in court is determined by its probative value. Probative value is a legal term that represents the extent to which a piece of evidence makes a proposition more or less likely to be true (Kaye, 1986). For instance, a suspect ID may make the proposition "the suspect is the perpetrator" more likely to be true. In this sense, the suspect ID is probative of guilt (from here on the term 'probative value' is shorthand for the probative value of guilt). The probative value of a suspect ID can be measured in a variety of ways. The most common measures of probative value derive closely from Bayes' Theorem (e.g. Bayes, 1763; Davis & Follette, 2002; Kaye & Koehler, 2003; Wells & Lindsay, 1980). Bayesian statistics have been argued to be well suited for a broad range of legal reasoning (e.g. Feinberg & Finklestein, 1996) and have been put to use to measure the probative value of legal evidence in a number of actual court cases (Finkelstein, 1978; Finkelstein & Fairley, 1970; Fienberg & Kadane, 1983; Goodman 1999; Marshall and Wise 1975; Satake & Amato 1999). However, another way to measure the reliability of a suspect ID is by conducting confidence-accuracy characteristic (CAC) analysis (Mickes, 2015). In fact, the results from CAC analysis may be more informative to triers of fact. Both methods are discussed in detail below.

Bayesian Statistics Approach

During a criminal trial, the suspect's guilt or innocence often cannot be definitively established. Because of this, a principal goal of the adjudicative process is to determine the probability of guilt, which can be denoted as P(G). This is referred to as the *prior* probability and is the probability of guilt prior to considering the suspect ID (Wells, 2003; Clark, 2012). In laboratory experiments, the prior probability is calculated by computing the proportion of lineups that contain the perpetrator (also referred to as the guilty suspect base rate). It is typically the case that half of the lineups contain the perpetrator and half of the lineups contain the innocent suspect meaning that P(G) is typically set to .50 (i.e. there is a 50% chance the suspect is guilty before knowing whether the participant identified the suspect). However, P(G) in the real world likely varies across police jurisdictions, investigators, and criminal cases (Clark, 2012). For instance, P(G) will be

quite high for a suspect whose DNA was collected from a crime scene (i.e. there is a high probability that the suspect is guilty), whereas P(G) will be much lower for a suspect who was arrested due to an anonymous tip (i.e. there is a low probability that the suspect is guilty). If an eyewitness identifies a suspect, then the trier of fact must update their belief in P(G). This is done by determining the probability of guilt given a suspect ID, denoted as P(G|ID). This is referred to as the *posterior* probability (Bayes, 1763).

Difference Measures of Probative Value

Friedman (1986) proposed that the greater the difference between the prior and the posterior probability, the greater the probative value of the evidence. This can be stated mathematically as:

$$Probative\ value = P(G|ID) - P(G) \quad (1)$$

This measure of probative value was directly motivated by the wording of the US Federal Rule of Evidence 401 which, at the time, defined relevant evidence as having "any tendency to make the existence of any fact that is of consequence to the determination of the action more probable or less probable than it would be without the evidence." Equation 1 can be considered as a mathematical restatement of rule 401. It indicates that a suspect ID has probative value if the posterior probability differs from the prior probability. If the suspect ID is unreliable, perhaps because the eyewitness is nearly blind, then the difference between the posterior and prior probability will be small. Whereas, if the suspect ID is reliable, then the posterior probability should be larger than the prior probability. It is ultimately up to the trier of fact to determine whether the suspect ID is reliable enough to warrant conviction. Although this measure satisfies rule 401, it is difficult to determine the probative value of suspect IDs using this measure because P(G) in real lineups is unknown (Ellman & Kaye, 1979; Aitken, 1995; Wells, 2003; Wixted & Mickes, 2012). Even if there was a way to accurately estimate P(G), many lawyers assume that estimating P(G) is within the province of the jury and should not be considered in court by forensic experts (Fenton & Neil, 2011).

Ratio Measures of Probative Value: The Diagnosticity Ratio

In eyewitness identification experiments, probative value is often calculated as a ratio of the correct ID rate to the false ID rate, commonly referred to as the diagnosticity ratio (DR) or the positive likelihood ratio (Kaye, 1986; Wells & Lindsay, 1980; Wells & Olson, 2002; Clark, 2012). The DR is stated mathematically as:

$$DR = \frac{Correct\ ID\ rate}{False\ ID\ rate} \tag{2}$$

The interpretation of the DR is straightforward: if, for example, the correct ID rate equals 0.8 and the false ID rate equals 0.2, the DR equals 4.0 (i.e. 0.8 / 0.2 = 4.0), meaning that a suspect ID is four times more likely when the suspect is guilty than when the suspect is innocent. As this ratio grows, it signals that a suspect ID is increasingly likely to occur when the suspect is guilty than when the suspect is innocent. If the ratio equals 1.0, the suspect ID has no probative value. If the ratio is less than 1.0, then a false ID is more likely than a correct ID. In this case, the suspect ID is probative of innocence rather than guilt. Because P(G) in real lineups remains unknown, the fact that the DR can provide a measure of probative value irrespective of P(G) is, in part, why eyewitness researchers (Wells & Lindsay, 1980) and other organizations such as the Royal Statistical Society (Puch-Solis, Roberts, Pope, & Aitken, 2012) have advocated its use in court.

Confidence and Accuracy Approach

Courts are often confronted with a situation where an eyewitness has identified the suspect with some degree of confidence (e.g. the eyewitness is 90% confident the suspect is the perpetrator) and there is very little or no other evidence available to help determine a verdict. It is estimated that each year in the US roughly 77,000 individuals become suspects in these types of cases (Goldstein, Chance, & Schneller, 1989). How can experimental studies of eyewitness identification help courts assess the probative value of suspect IDs made with some degree of confidence?

One way to try to help the court would be to compute the DR from laboratory experiments (Wells & Lindsay, 1980). It is possible to compute the DR across degrees of

confidence (e.g. Brewer & Wells, 2006; Lindsay & Wells, 1985), but the question triers of fact have about a testifying eyewitness who has identified a suspect is: How *accurate* are suspect IDs given the level of confidence expressed by an eyewitness? Two analyses aim to answer this question by measuring the relationship between confidence and accuracy.

Calibration Analysis

Although results from early experimental studies have been interpreted to mean that eyewitness confidence is an unreliable indicator of suspect ID accuracy, confidence collected at the time of the initial suspect ID is, in fact, a strong predictor of suspect ID accuracy (Wixted, Mickes, Clark, Gronlund, & Roediger, 2015). This confidence-accuracy (CA) relationship is typically strong for eyewitnesses who make a suspect ID in the laboratory (Horry, Palmer, & Brewer, 2012; Sauer, Brewer, Zweck, & Weber, 2010) and in the field (Behrman & Davey, 2001; Wixted, Mickes, Dunn, Clark, & Wells, 2016). Juslin, Olsson, and Winman (1996) measured the CA relationship by conducting calibration analysis. This approach measures accuracy for each level of confidence as:

$$A = \frac{\#\ Correct\ IDs}{(\#\ Correct\ IDs\ +\ \#\ False\ IDs\ +\ \#\ FillerIDs)} \tag{3}$$

, where *A* is the accuracy of eyewitness IDs. This equation counts the number of correct IDs and divides that number by the total number of IDs (i.e. by adding the number of correct IDs, false IDs, and filler IDs) for a particular level of confidence. For example, accuracy for high levels of confidence would be:

$$A_{high} = \frac{\#\ Correct\ IDs_{high}}{\#\ Total\ IDs_{high}} \tag{4}$$

Calibration analysis measures how well eyewitnesses can calibrate their subjective confidence in their identification decision with their objective accuracy in those identifications. An eyewitness is said to be "over-confident" if their confidence exceeds their accuracy and "under-confident" if their accuracy exceeds their confidence. Perfect calibration occurs when an eyewitness's reported confidence matches their accuracy. For

example, if an eyewitness is 70% confident that the suspect is the perpetrator, perfect calibration would mean that the eyewitness is correct 70% of the time.

Confidence-Accuracy Characteristic Analysis

Confidence-accuracy characteristic (CAC) analysis measures the CA relationship in a similar way that is *more* informative to the trier of fact (see Chapter 2 for more details; Mickes, 2015). A CAC analysis plots suspect ID accuracy across levels of confidence. Suspect ID accuracy, denoted as *A*, is calculated as:

$$A = \frac{\#\ Correct\ IDs}{(\#\ Correct\ IDs\ +\ \#\ False\ IDs)} \quad (5)$$

, where *A* equals the number of correct IDs divided by the total number of suspect IDs (i.e. by adding the number of correct and false IDs). Crucially, filler IDs are not included in this measurement because triers of fact are mainly concerned with the accuracy for suspects. Of course, the number of correct and false IDs obtained in an experiment might fluctuate depending on the number of times the perpetrator is present in the lineup. That is, suspect ID accuracy might fluctuate depending on P(G). A low P(G) could yield few correct IDs and possibly many false IDs (because the perpetrator is rarely present in the lineup), whereas a high P(G) could yield many correct IDs and few false IDs (because the perpetrator is often present in the lineup). Thus, it is important to consider how suspect ID accuracy varies across a range of P(G) values.

The results from a recent police department field study (Wixted et al., 2016) estimate suspect ID accuracy to be roughly 97% for actual eyewitnesses who were highly confident at the time of their initial identification from a lineup. Remarkably, this estimate did not vary significantly across P(G) values .25, .50, and .75. These results are very similar to the results often found in the laboratory. Wixted, Read, Lindsay, & Columbia (2016) reanalyzed data from several eyewitness identification experiments which had P(G) values .50 or .75. Across these studies, suspect ID accuracy was roughly 97% when participants were 90 – 100% confident. Together, these results make it clear that a suspect identified by an eyewitness with high confidence has a high likelihood of being the person

who actually committed the crime (i.e., a high likelihood of being guilty), and this is true when the prior probability is as high as .75 and when the prior probability is as low as .25.

The Diagnosticity Ratio or Confidence-Accuracy Characteristic Analysis

Both the DR and CAC analysis effectively capture the probative value of suspect IDs, but when courts need to know the *accuracy* of suspect IDs made with a particular level of confidence (as is often the case in the US), CAC analysis seems generally preferable (Mickes, 2015). To demonstrate this point, consider the case Jackson v. Fogg (1978). In June 1970, an armed man entered a diner in Queens County, New York and announced, "This is a stickup." Hearing these words, customers and staff members attempted to quickly find cover. The gunman then fired a single shot, mortally wounding the bartender. New York police detectives investigating the shooting were unable to locate any tangible evidence pertaining to the identity of the perpetrator, but four eyewitnesses identified Edmond Jackson who had been brought to the police station on an unrelated issue. At trial, no other evidence was brought forth connecting Jackson to the crime and the jury found Jackson guilty. He was sentenced to 20 years to life in prison. Several years later, the conviction was reversed in part because each of the four eyewitnesses had seen the perpetrator for only a few seconds and, thus, did not have a good opportunity to view the perpetrator's face.

Are these Suspect IDs less Reliable?

The court in this case is concerned with whether the suspect IDs were less reliable because the perpetrator's face was shown for just a short amount of time. When examining the effect of exposure duration on eyewitness reliability, and when confidence ratings are recorded, one can easily plot CACs and DRs. Take, for example, data from experiment 1 in Palmer, Brewer, Weber, and Nagesh (2013). Participants viewed a perpetrator for 5 (short exposure) or 90 seconds (long exposure) before attempting to identify the perpetrator from a perpetrator-present or perpetrator-absent lineup. During the lineup, participants indicated how confident they were in their identification decision using an 11-point confidence scale (ranging from 0% = just guessing to 100% = absolutely certain).

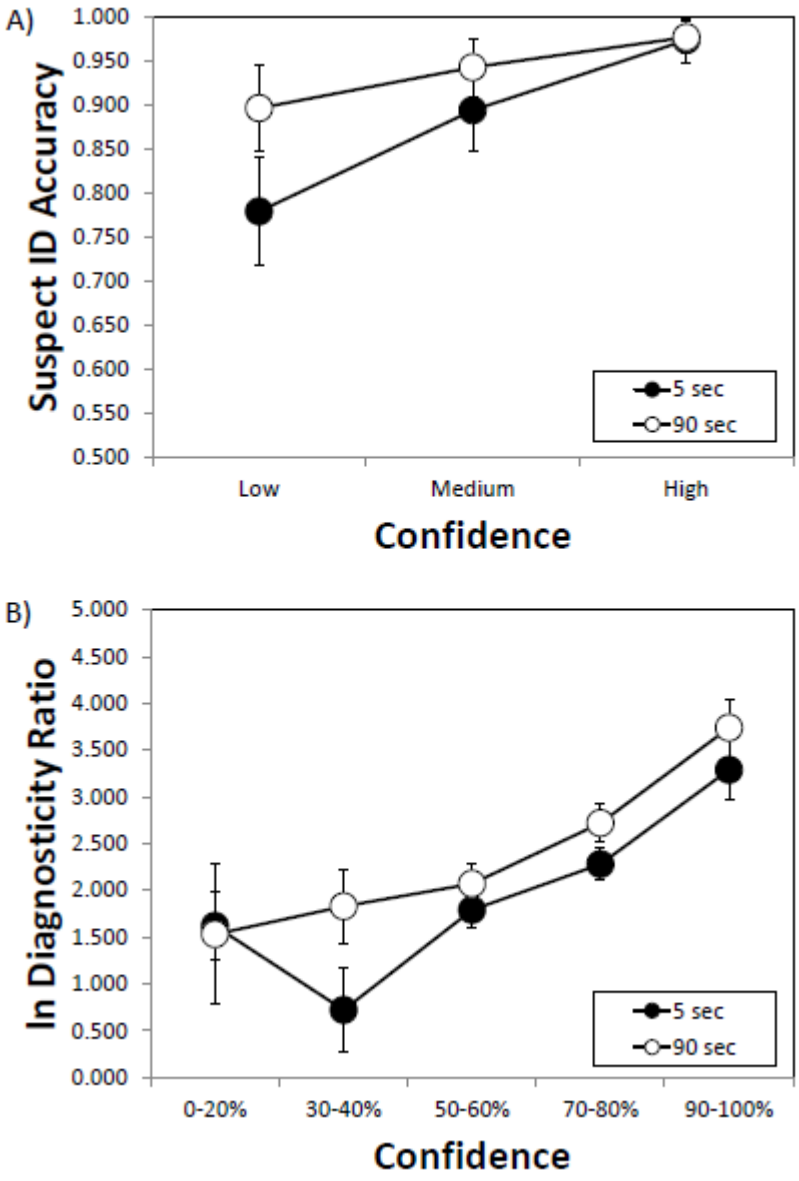

Figure 1. The reliability of a suspect ID can be measured using the DR or by measuring the CA relationship. Figure 1A plots the relationship between confidence and suspect ID accuracy, whereas Figure 1B plots the relationship between confidence and the ln DR. The bars are the standard error for each level of confidence.

Figure 1A plots the CACs for short (5 seconds) and long (90 seconds) exposure conditions. Figure 1B plots the natural log (ln) of the DR across each level of confidence for short and long exposure conditions. Note that the ln DR is conceptually similar to the

traditional DR (i.e. the higher values correspond to greater probative value). First, both figures clearly show that a suspect identified by an eyewitness with high confidence, as opposed to low confidence, is more likely to be the person who actually committed the crime. As Palmer et al. state, "In all conditions, the accuracy and informativeness [i.e. probative value] of positive identifications increased with confidence, especially toward the upper end of the confidence scale" (p. 62-63). Second, the experimental manipulation of exposure duration had no meaningful effect on the probative value of suspect IDs. Both figures show no significant differences between short and long exposure duration on the reliability of suspect IDs across medium (70-80%) and high (90-100%) levels of confidence. Based on these results, the fact that the four eyewitnesses only saw the perpetrator's face for a few seconds does *not* indicate that their suspect IDs were less reliable.

Yet, CAC analysis makes this point more clearly. For suspect IDs made with high confidence, for example, suspect ID accuracy is 98% for *both* short and long exposure conditions. This means that participants who saw the perpetrator for only a few seconds and confidently identified the suspect from a lineup are no more likely to make a false ID than participants who saw the perpetrator for a longer duration. In both conditions, there is a 98% chance of a correct ID and just a 2% chance of a false ID. The ln DR calculated from the same dataset for suspect IDs made with 90-100% confidence is 3.74 for the long exposure condition and 3.29 for the short exposure condition. This is not a statistically significant difference, which is informative to the trier of fact, but because the ln DR is not a measure of proportion correct (i.e. is not a measure of *accuracy*), triers of fact are not privy to the error rate. Thus, when measuring the reliability of suspect IDs, CAC analysis seems more informative to the trier of fact than the DR.

Assisting the Criminal Justice System: Consultants to Policymakers

In the early-1970s eyewitness researchers began to provide expert testimony in court with the aim of spotting mistakes in eyewitness testimony. Several years later, in the late-1970s, eyewitness researchers aimed to prevent eyewitness errors from occurring in the first place (Wells, 1978). This new perspective held that some variables, known as

system variables, affect eyewitness memory and are under the control of the criminal justice system, whereas other variables, known as estimator variables, affect eyewitness memory, but are beyond the control of the criminal justice system. Thus, ways of interviewing an eyewitness (e.g. by using the Cognitive Interview; Fisher & Geiselman, 1992) and ways of collecting identification evidence (e.g. by using the sequential lineup procedure; Lindsay & Wells, 1985) are system variables because the criminal justice system controls these procedures. Factors that are beyond the control of the criminal justice system that may affect eyewitness memory are estimator variables. Laboratory experiments indicate that the amount of time the perpetrator was in view (e.g. Memon, Hope, & Bull, 2003), the distinctiveness of the perpetrator's face (e.g. Light et al. 1979), and the presence of a weapon (i.e. by drawing attention away from the perpetrator's face; Loftus et al., 1987; Steblay, 1992) may impact eyewitness memory. Each of these factors are estimator variables because the criminal justice system can, at best, *estimate* the extent to which these factors affect eyewitness memory.

The value of distinguishing between system and estimator variables is that it clearly separates variables that can be reformed by the criminal justice system (i.e. system variables) from variables that cannot be reformed (i.e. estimator variables). That is, although estimator variables may be manipulated in laboratory experiments, they are uncontrollable in actual criminal situations (Cutler, Penrod, & Martens, 1987). Whereas, if current police procedures are substandard, system variable research can inform policymakers of alternative procedures to help improve eyewitness memory.

Review of System Variable Research

Research on system variables increased in the following decades and several reforms were introduced (Gronlund, Mickes, Wixted, & Clark, 2015); but simply having a strong research base was not sufficient for policymakers to take notice and adopt these reforms (Wells et al., 2000). Eyewitness researchers put some pressure on policymakers to adopt these reforms by scrutinizing faulty police procedures in court. In the US, courts could dismiss eyewitness evidence if it was collected under biased or risky circumstances, but the courts have been reluctant to do this. The courts have instead preferred to let the

trier of fact evaluate the reliability of the eyewitness evidence even if it was collected under extremely biased circumstances (Loftus & Doyle, 1997). The catalyst for reform has been the staggering number of wrongfully convicted individuals originally convicted due to false IDs and later exonerated based on DNA evidence.

The Innocence Project, whose mission it is to lobby on behalf of wrongfully convicted individuals, has helped release 349 innocent suspects (as of 07/02/2017) through the use of DNA testing. Of those, roughly 72% were convicted either solely or in large-part due to false IDs (Innocence Project, 2017). These DNA-based exonerations encouraged U.S. Attorney General Janet Reno to order a panel to discuss and develop national guidelines on the appropriate methods for collecting eyewitness identification evidence (Wells et al., 2000). In October 1999, the U.S. Department of Justice released a document entitled *Eyewitness Evidence: A Guide for Law Enforcement* (Technical Working Group for Eyewitness Evidence, 1999), which established a host of guidelines that aimed to improve eyewitness memory (Gronlund et al., 2015). The reforms focused on improving fundamental aspects of the identification procedure such as the wording of pre-lineup instructions, the selection of fillers, and the way lineups are constructed (Clark, 2012).

Misuse of the Diagnosticity Ratio

Although heralded as a "successful application of eyewitness research", "from the lab to the police station" (Wells et al., 2000), many of the reforms have either failed to improve eyewitness memory or have made it worse (Gronlund et al., 2015). This undesirable consequence occurred in large part because the wrong statistic – the DR – was used as the basis for many of the recommended reforms (Gronlund et al., 2015). Crucially, the sequential lineup procedure was included in the guidelines as a recommended method for constructing a lineup almost entirely on the basis that the DR was higher for the sequential lineup procedure than the traditionally used simultaneous lineup procedure (Steblay et al., 2001). The better lineup procedure was thought to be the one that resulted in more reliable (more probative) suspect IDs (as indicated by the DR). To illustrate why the DR should not have been used as the basis for these policy recommendations, the

debate between the traditional simultaneous lineup procedure and its recommended replacement, the sequential lineup procedure, will be discussed. The point of this discussion is to show that the DR should not serve as the basis for policy recommendations when the goal is to decrease the number of false IDs and misses.

Simultaneous vs. Sequential Lineups

The simultaneous lineup procedure is the most commonly constructed lineup in the United States (Police Executive Research Forum, 2013). The lineup consists of six photos presented simultaneously to the eyewitness in a 3 X 2 matrix. The eyewitness views all of the six photos and either chooses an individual from the lineup as the perpetrator or rejects the entire lineup in the event the perpetrator is absent. In attempts to improve the lineup, Lindsay & Wells (1985) developed the sequential lineup procedure. The sequential lineup procedure consists of photos of individuals presented one at a time (i.e. in sequence) to an eyewitness. For each photo, the eyewitness declares whether the individual in the photo is or is not the perpetrator. In the laboratory, researchers recommend terminating the procedure after making an identification, but, in practice, the procedure typically terminates after the eyewitness has seen all of the lineup members.

Researchers determine the performance of a lineup procedure by measuring the correct ID rate and false ID rate. When comparing two lineup procedures such as procedure A and procedure B, for example, procedure A is the superior procedure if procedure A yields a lower false ID rate and higher (or at least the same) correct ID rate. Alternatively, procedure A may yield a higher correct ID rate and lower (or at least the same) false ID rate and be the superior procedure. However, when comparing simultaneous and sequential lineups, a more ambiguous outcome arises (Clark, 2012). Meta-analyses show that the sequential lineup procedure often yields a lower false ID rate and a slightly lower correct ID rate than the simultaneous lineup procedure (Steblay, Dysart, Fulero, & Lindsay, 2001; Steblay, Dysart, & Wells, 2011). Note that with this pattern, superiority cannot be determined simply by comparing the overall correct and false ID rates from the two procedures (Mickes et al., 2012). Yet, the sequential lineup procedure was still deemed the superior procedure (Steblay et al., 2011).

Sequential Superiority Effect

Eyewitness identification researchers argue that the sequential lineup procedure is superior largely because it yields a higher DR (Steblay et al., 2011; although see Gronlund, Carlson, Dailey, & Goodsell, 2009). If the sequential procedure, despite having a lower correct ID rate and lower false ID rate, has a higher DR, then the decrease in false ID rate is proportionally greater than the corresponding decrease in correct ID rate. In other words, some researchers claim that the benefits gained by using the sequential procedure (the large decrease in false ID rate) outweigh the negatives (the slight decrease in correct ID rate). For this reason, the sequential procedure has been dubbed the "superior procedure" (Steblay, et al. 2011). Based on this logic, the sequential procedure was included in the national guidelines (Technical Working Group, 1999) and approximately 30% of jurisdictions throughout the US have since switched to the sequential procedure (Police Executive Research Forum, 2013).

Discriminability

However, the procedure that maximizes *discriminability* – the ability to distinguish perpetrators from innocent suspects – is the superior procedure (Mickes et al., 2012). The highest level of discriminability occurs when perpetrators are always identified and innocent suspects are never identified (i.e. correct ID rate = 1.0, false ID rate = 0.0). The lowest level of discriminability occurs when perpetrators are identified as often as innocent suspects (i.e. correct ID rate = false ID rate). A recent US National Academy of Sciences (NAS) committee commissioned to assess the current state of eyewitness identification research concluded that "there should be no debate about the value of greater discriminability – to promote a lineup that yields less discriminability would be akin to advocating that the lineup be performed in dim instead of bright light" (National Research Council, 2014, p. 80). Thus, the lineup procedure that yields greater discriminability is the superior procedure.

Response Bias

Can the DR, a measure of probative value for suspect IDs (e.g. Wells & Lindsay, 1980), measure discriminability? No. This is because the DR conflates discriminability with the eyewitness' likelihood to choose, or not choose, a member from the lineup, commonly referred to as *response bias* (Wixted & Mickes, 2012). Eyewitnesses may be more or less likely to choose someone from the lineup (i.e. more or less biased to choose), but may be equally proficient at discriminating perpetrators from innocent suspects (i.e. have the same discriminability). Because the DR cannot properly separate measurements of response bias from measurements of discriminability, the DR cannot determine whether the sequential procedure increases discriminability or simply increases response bias. Some have argued that the reason the DR is higher for the sequential procedure is precisely because the sequential procedure induces conservative responding (e.g. Gronlund et al., 2009; Meissner, Tredoux, Parker, & Maclin, 2005; Mickes et al., 2012). For this reason the NAS committee has advocated the use of other methods of measuring discriminability instead of the DR. One such method endorsed by the committee is receiver operating (ROC) analysis which the committee concluded as, "…a positive and promising step, with numerous advantages" (National Research Council, 2014, p. 59). The main advantage being that differences in discriminability and response bias are apparent in ROC curves (Green & Swets, 1966; Macmillan & Creelman, 2005; Wixted & Mickes, 2012; Gronlund, Wixted, & Mickes 2014; National Research Council, 2014).

Receiver Operating Characteristic Analysis

A comparison between the overall correct and false ID rates for the simultaneous and sequential lineup procedure has repeatedly been the basis for determining lineup superiority, often calculated in the form of a DR (e.g. Steblay et al., 2001; 2011). However, in order to determine which procedure yields greater discriminability, the correct and false ID rates for identifications made with increasingly conservative responding must be compared as well. These correct and false ID rates can be obtained from instruction biasing conditions (e.g. Mickes et al., in submission) or can be obtained, more conveniently, by collecting confidence in the identification decision (e.g. Seale-Carlisle &

Mickes, 2016). ROC analysis plots the correct and false ID rate for each level of confidence, from highly confident identifications (i.e. very conservative identifications) to identifications made with lower confidence (i.e. increasingly liberal identifications). This is how ROC plots have typically been constructed in the eyewitness identification field (e.g. Mickes et al., 2012) and confidence-based ROC analysis is typical in the fields of experimental psychology and radiology (Wixted & Mickes, 2012). By using confidence judgments to construct the entire ROC, the comparison between the simultaneous and sequential lineup procedure can be made on the basis of the family of correct and false ID rates, rather than on the basis of a single, overall correct and false ID rate.

Figure 2 shows the overall correct and false ID rate for the simultaneous lineup and the overall correct and false ID rate for the sequential lineup. This data is taken from Table 3 of the meta-analysis reported by Steblay et al. (2011). It is clear that the sequential lineup yields a lower correct ID rate and a lower false ID rate than the simultaneous lineup. However, once the family of correct and false ID rates has been plotted (so that the entire ROC is constructed for the simultaneous and sequential lineup procedures), three patterns may emerge (Mickes et al., 2012). First, the points may fall on the same ROC (Figure 2A), in which case the data would show that the two lineup procedures yield the same discriminability, and the sequential lineup procedure simply induces more conservative responding (Ebbesen & Flowe, 2002; Gronlund et al., 2009; Meissner et al., 2005). Second, the sequential lineup ROC may fall above the simultaneous lineup ROC (Figure 2B). This pattern would indicate that the sequential lineup procedure yields greater discriminability and is, in fact, the superior procedure – for any given false ID rate, the sequential lineup would yield a higher correct ID rate. A third possibility is that the simultaneous lineup ROC falls above the sequential lineup ROC (Figure 2C). This pattern would indicate the exact opposite of the "sequential superiority" claim as the data would show the simultaneous lineup procedure to be superior.

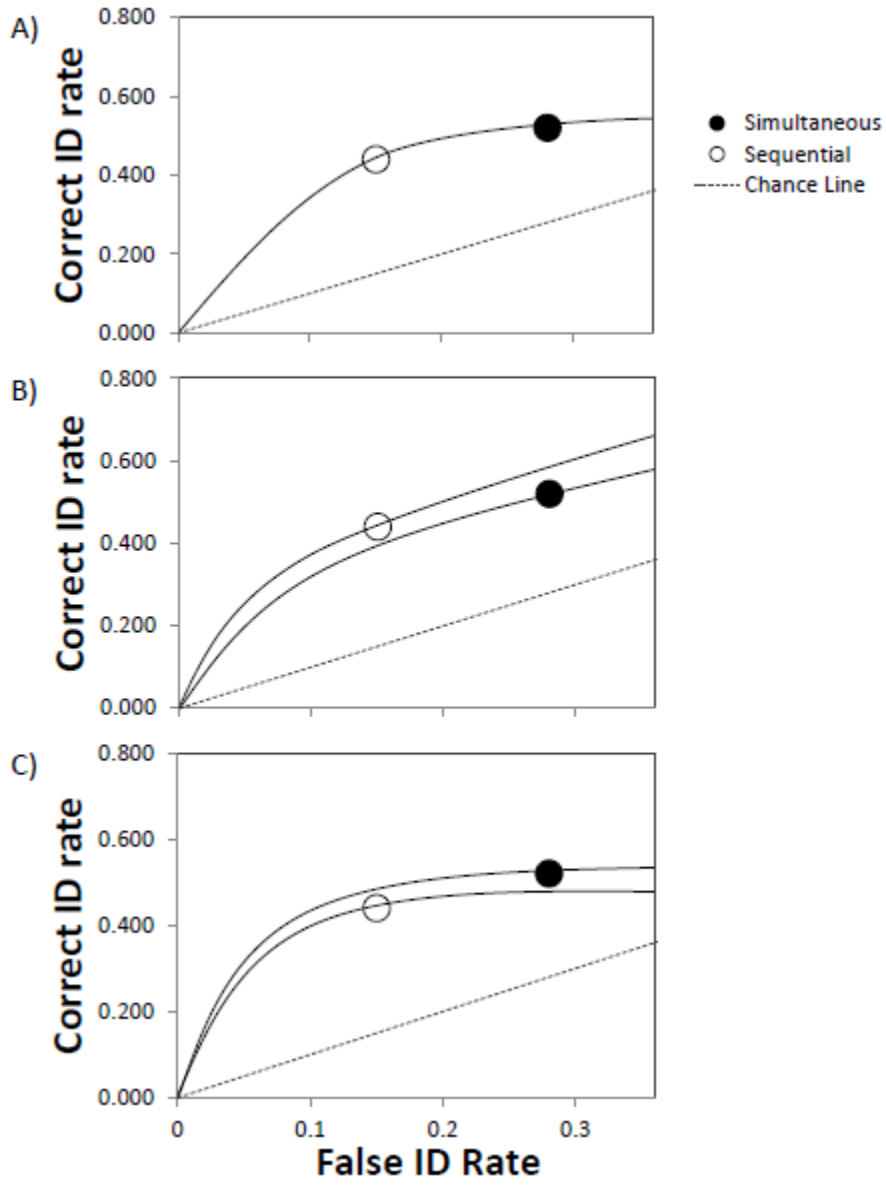

Figure 2. The data points in this figure are from Table 3 of the meta-analysis reported by Steblay et al. (2011). Three hypothetical ROC curves are shown in this figure. If there is no significant difference in discriminability between simultaneous and sequential lineups, then there should be one ROC as shown in Figure 2A. If the sequential lineup yields greater discriminability, then the sequential lineup ROC should be higher than the simultaneous ROC, as shown in Figure 2B. If the simultaneous lineup yields greater discriminability, then the simultaneous lineup ROC should be higher than the sequential lineup ROC, as shown in Figure 2C. This point has been expressed in Mickes et al. (2012).

Simultaneous vs. Sequential Lineups

Mickes et al. (2012) conducted ROC analysis to determine whether the sequential procedure yields greater discriminability than the simultaneous procedure. They had participants study a video of a mock crime and attempt to identify the perpetrator from either a 6-person simultaneous lineup or a 6-person sequential lineup. In order to construct the ROC, they had participants report their confidence in their identification decision using an eleven point confidence rating scale (0% = just guessing to 100% = absolutely certain). The correct and false ID rate for each level of confidence, from highly confident identifications (i.e. very conservative identifications) to identification decisions made with lower confidence (i.e. increasingly liberal identifications), were plotted in ROC space (see Figure 3). Note that the x-axis, ranging from 0 – .08, is much shorter than the y-axis, ranging from 0 – .80, because the false ID rates were estimated by dividing the number of filler IDs by the size of the lineup, which produces very small false ID rates (see Chapter 2 for further discussion).

Mickes et al. (2012) found that 1) the simultaneous lineup yields a higher ROC curve and, thus, greater discriminability than the sequential lineup and 2) the sequential lineup induces more conservative responding than the simultaneous lineup. This result has, thus far, been replicated by four independent laboratories (e.g. Anderson, Carlson, Carlson, & Gronlund, 2014; Carlson & Carlson, 2014; Dobolyi & Dodson, 2013; Gronlund et al., 2012) and data from two field studies provide converging evidence (Amendola & Wixted, 2014; Wixted et al., 2016). Thus, when ROC analysis is conducted, the superior procedure is the simultaneous procedure rather than the sequential procedure. Police jurisdictions that have switched to the sequential procedure have, in fact, made it harder for eyewitnesses to discriminate the guilty suspect from the innocent suspect. When the goal is to increase eyewitnesses' ability to discriminate the innocent suspect from the guilty suspect, policymakers should adopt the procedure that yields the greatest ROC rather than the procedure that yields the greatest DR.

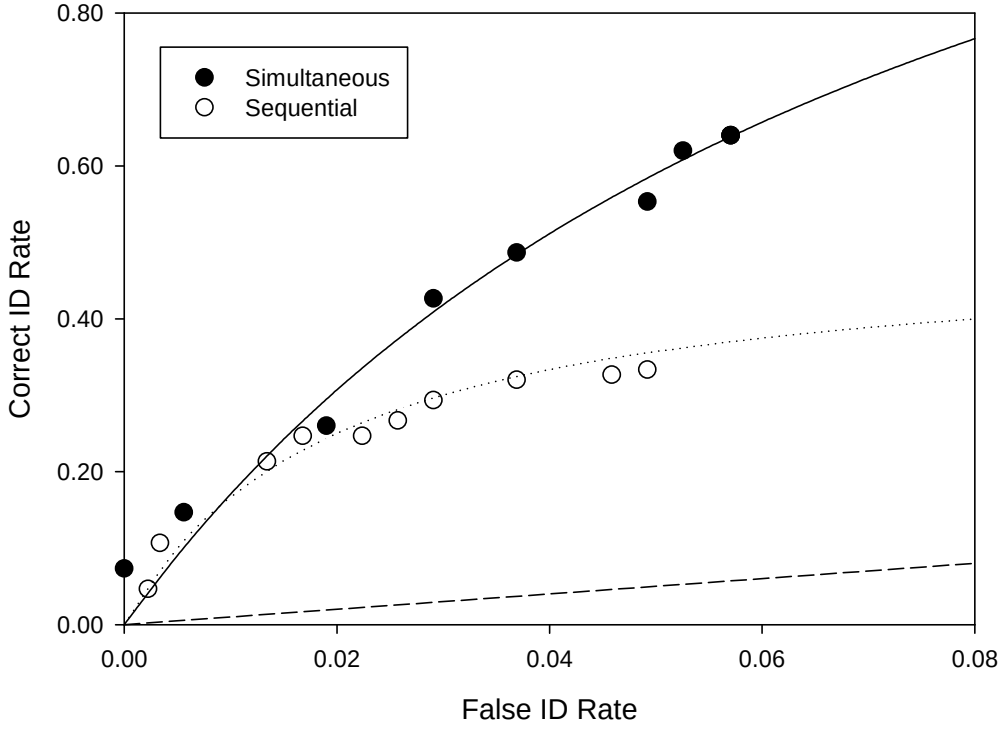

Figure 3. Receiver operating characteristic (ROC) curves for simultaneous and sequential lineup procedures in Experiment 1a of Mickes, Flowe & Wixted (2012). The dashed line represents chance performance.

Signal-Detection Theory

Why is discriminability higher for the simultaneous lineup? Because theories of discriminability in the domain of eyewitness identification were nonexistent, Wixted and Mickes (2014) extended a signal-detection-based model that has served as a theoretical bedrock for recognition decisions since the 1950s (Egan, 1958), and applied it to eyewitness identifications from police-constructed lineups. This model is composed of two parts: a general signal-detection based model of eyewitness identification and a specific *diagnostic feature-detection* hypothesis which predicts greater discriminability for the simultaneous lineup. The basic tenants of signal-detection theory and its application as a model of recognition memory are first reviewed in order to discuss the specifics of the diagnostic feature-detection hypothesis.

Standard Recognition Memory Experiment

The standard recognition memory experiment consists of two phases: a study phase and a test phase. During the study phase, participants are presented a list of items to memorize (e.g., a list of words or a list of faces); these items are called "targets." During the test phase, participants are asked to distinguish between targets and new items that were not studied called "lures." The targets are randomly intermixed with the lures and are presented one at a time for an "old" or "new" decision. A correct response for a target is old; these responses are called "hits." A correct response for a lure is new; these responses are called "correct rejections." Incorrectly declaring a target as new is a "miss" and incorrectly declaring a lure as old is a "false alarm." Table 2 shows these four decision outcomes.

Table 2

The four decision outcomes of a recognition memory experiment.

	Old	New
Target	Hit	Miss
Lure	False Alarm	Correct Rejection

The proportion of targets correctly identified as old is the hit rate and the proportion of lures incorrectly identified as old is the false alarm rate. Hit and false alarm rates are computed for each participant. A participant might, for example, correctly declare 80 out of 100 targets as old (i.e. hit rate equals 80 percent, miss rate equals 20 percent) and might incorrectly declare 30 out of 100 lures as old (i.e. false alarm rate equals 30 percent, correct rejection rate equals 70 percent).

Signal-detection theory is capable of providing an interpretation of hit and false alarm rates obtained from recognition memory experiments (Egan, 1958; Green & Swets, 1966). The standard, prototypical signal-detection model is illustrated in Figure 4. According to this view, recognition decisions are based on the strength of a memory signal in relation to a decision criterion. This approach assumes two Gaussian distributions: one

target distribution and one lure distribution. The difference in strengths between the two distributions reflects the amount of strength added to target items on account of being presented on the study list; lure items receive no added strength as they were not presented on the list but are noisy, hence a Gaussian distribution. In a recognition task, any item that generates a memory strength signal exceeding the criterion is declared to be old (i.e. the participant recognized the item). If the item fails to generate enough strength to surpass the criterion, the item is declared new (i.e. the participant did not recognize the item). The light grey area of the target distribution to the right of the decision criterion represents the hit rate, whereas the dark grey area of the lure distribution to the right of the decision criterion represents the false alarm rate.

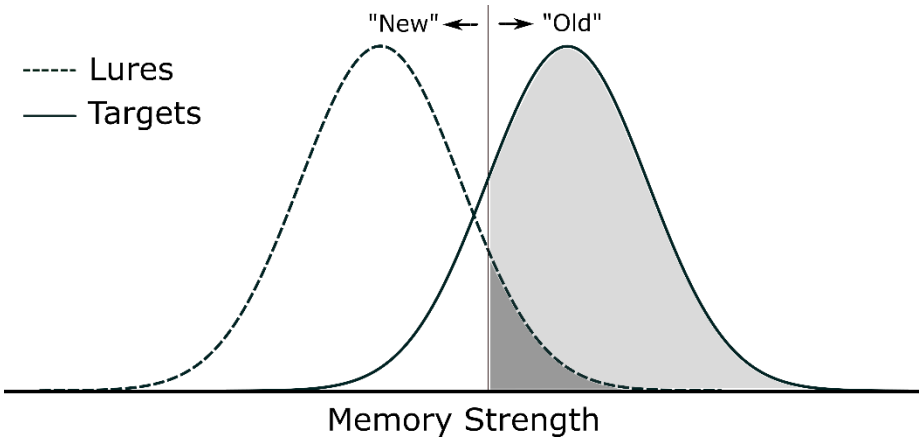

Figure 4. The standard signal detection model consists of two distributions: one lure distribution and one target distribution. These distributions rest along a memory strength axis. A decision criterion is placed somewhere along the axis such that memory signals for items exceeding that criterion are recognized as old and memory signals for items that fail to reach that criterion are determined to be new.

Response Bias

Response bias refers to the likelihood of declaring an item as old (a bias to choose an item as having been recognized) and this likelihood is represented by the location of the decision criterion along the memory strength axis. Shifting the decision criterion to the right means that items will have a lower likelihood of being declared old (Figure 5A). Only items that generate a very strong memory signal will be declared old; most items

will be declared new. As such, the hit rate and false alarm rate will decrease as a smaller portion of the target and lure distributions fall to the right of the decision criterion (i.e. the shaded regions of the distributions are smaller). Shifting the decision criterion to the left (Figure 5B) means that items will have a greater likelihood of being declared old as many targets and lures will generate a strong enough signal to exceed the decision criterion. When this happens, the hit rate and false alarm rate will increase as a greater portion of the target and lure distributions fall to the right of the decision criterion (i.e. the shaded regions of the distributions are larger).

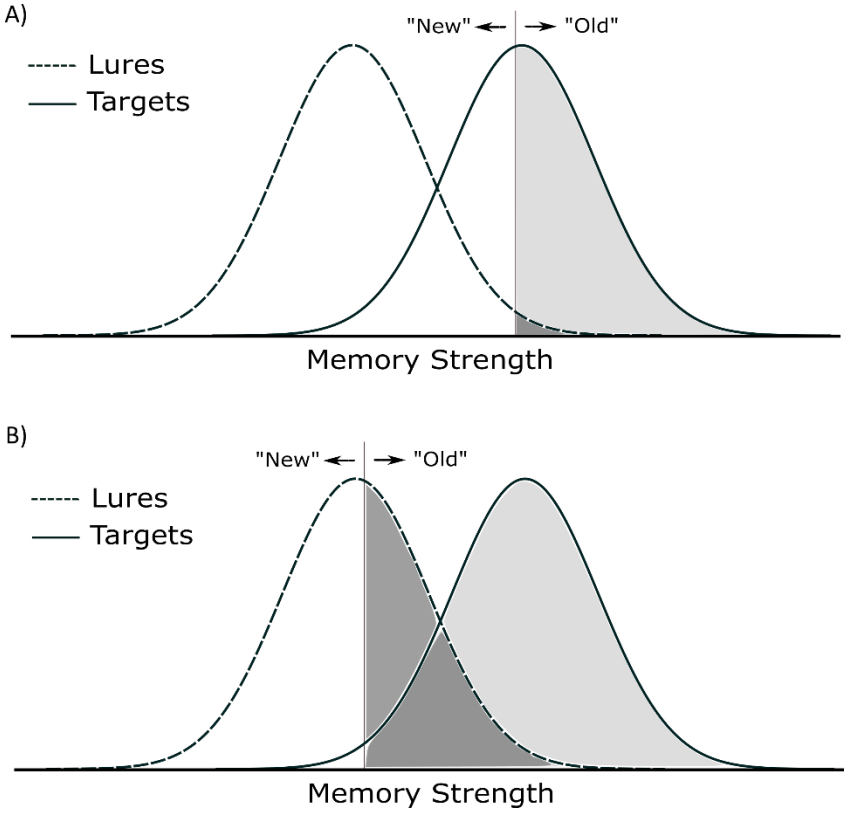

Figure 5. Different response biases. The location of the decision criterion can shift along the memory strength axis which can affect the hit and false alarm rates, but underlying discriminability does not change. Figure 5A represents conservative responding, whereas Figure 5B represents liberal responding.

Discriminability

The target and lure distributions can also shift along the memory strength axis. The extent to which the two distributions overlap reflects the participant's ability to discriminate targets from lures. High discriminability is evident when the target distribution is largely separated from the lure distribution (Figure 6A). The highest level of discriminability is when a target is always recognized and a lure is never recognized (i.e. a hit rate of 1.0 and a false alarm rate of 0). Low discriminability is evident when the target and lure distributions largely overlap (Figure 6B). The lowest level of discriminability is when targets and lures are recognized equally. In Figure 5 we can see that despite placing the decision criterion at different points along the memory strength axis (i.e. despite differences in response bias), discriminability is the same because the target and lure distributions have the same amount of overlap.

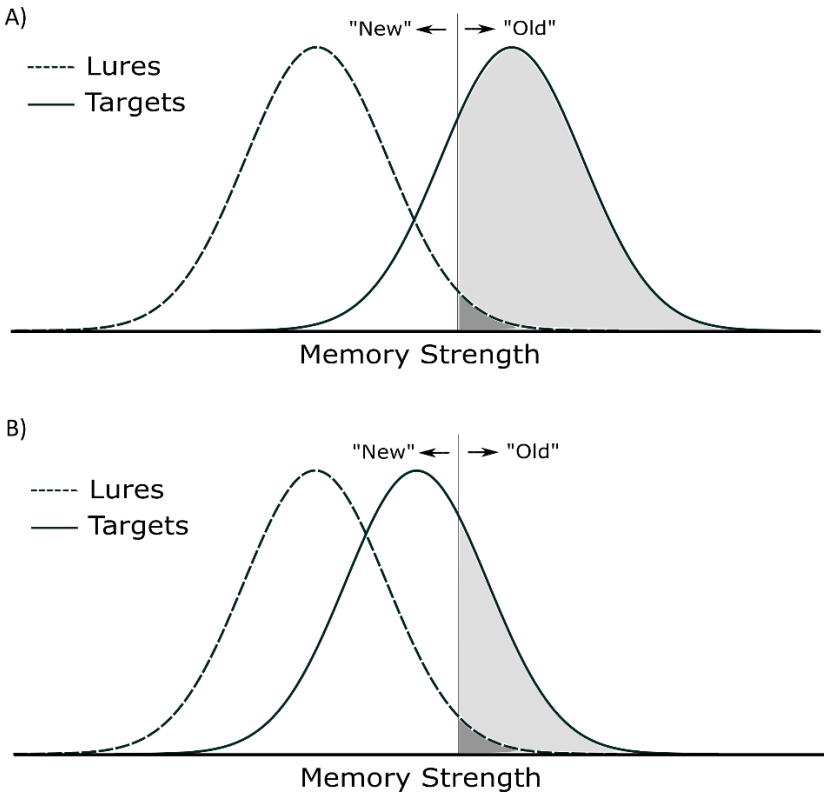

Figure 6. Different discriminability. The lure and target distributions can shift along the memory strength axis and the distance between these distributions depicts discriminability.

Signal-Detection Model and Lineups

The extension of the prototypical signal-detection model to police-constructed lineups is rather straightforward. A lineup consists of a police suspect (who may be innocent or guilty) and several fillers. The police should select fillers that match the general description of the perpetrator. A lineup is considered "fair", if an innocent suspect does not resemble the perpetrator more so than the other fillers (to ensure that the innocent suspect does not stand out). In this case, the innocent suspect distribution and the filler distribution are one and the same (and will from here on just be referred to as the innocent suspect distribution). Thus, there exists two distributions in this model: an innocent suspect distribution and a guilty suspect distribution (Figure 7). The amount of overlap between the innocent suspect distribution and the guilty suspect distribution represents the eyewitnesses' discriminability, and an identification is made only if the memory strength of the *most familiar* face in the lineup exceeds a decision criterion. The area of the guilty suspect distribution to the right of the decision criterion is the correct ID rate (shaded in light grey), whereas the area of the innocent suspect distribution to the right of the decision criterion is the false ID rate (shaded in dark grey). Although signal-detection models have been traditionally applied to standard list-learning recognition paradigms, this basic signal-detection model has been shown to fit experimental lineup data (e.g. Wixted et al., 2016; Seale-Carlisle & Mickes, 2016).

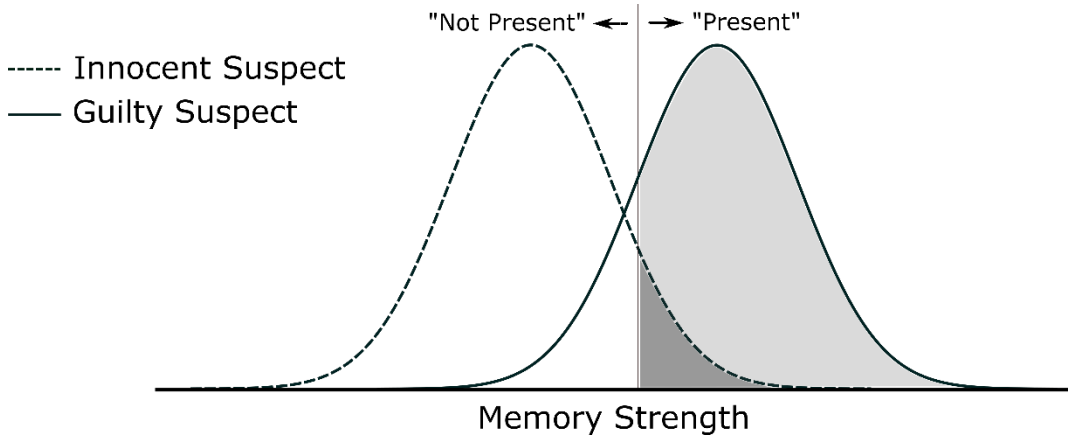

Figure 7. Signal-detection model for lineups consists of two distributions: one distribution for innocent suspects (and fillers for a fair lineup) and one distribution for guilty suspects.

Diagnostic-Feature-Detection Hypothesis

The *diagnostic-feature-detection* (DFD) hypothesis holds that the simultaneous presentation of faces better enables the eyewitness to detect the features that are unique to the perpetrator (i.e. features that are diagnostic of guilt) while discounting the features that the perpetrator shares with other members in the lineup (i.e. the features that are non-diagnostic; Wixted & Mickes, 2014). By focusing on the diagnostic features, eyewitnesses are able to improve their ability to discriminate innocent from guilty suspects. For example, if the perpetrator was a young, White male, then attaching weight to those features would not be helpful and would instead serve to impair discriminability because all of the lineup members would be young, White males. Having the faces presented simultaneously allows eyewitnesses to immediately detect and discount non-diagnostic features (e.g. the age, ethnicity, and gender of the suspect) and to instead attach more weight to features that are not shared and are thus more diagnostic. The sequential presentation of faces, on the other hand, limits the ability of the eyewitness to detect (and then discount) the common, non-diagnostic facial features, which means that eyewitnesses will tend to attach weight to diagnostic *and* non-diagnostic features, thereby reducing discriminability. These are currently untested predictions of the DFD hypothesis, but the research discussed in Chapter 3 and Chapter 4 are consistent with these predictions.

Preview of Upcoming Chapters

Chapter 2 contains the methodological details shared among the eight eyewitness identification experiments discussed in this thesis. In Chapter 3, a direct comparison between the US lineup and the UK lineup is made. In Chapter 4, the validity of the verbal overshadowing effect (Schooler & Engstler-Schooler, 1990) is re-examined and a diagnostic feature-detection account of the verbal overshadowing effect is proposed. In Chapter 5, the validity of the weapon focus effect (Loftus et al., 1987) is re-examined. In Chapter 6, several potential markers of eyewitness identification accuracy are compared to determine which marker is best. Chapter 7 discusses the general findings from these experiments as well as any overarching strengths and limitations of the research conducted.

Chapter 2

The methods that apply to the experiments in this thesis are described in this chapter.

General Procedure

The general procedure used in the experiments reported in this thesis consisted of two phases: a study phase and a test phase. During the study phase, participants took on the role of an eyewitness by watching a video of a mock crime. Each mock crime video consisted of a single perpetrator, though in some of the videos a victim was also present. After witnessing the crime, participants engaged in a distractor task. The distractor task was designed to limit the participants' ability to rehearse information from the video so that participants rely on their long-term memory (rather than short-term or working memory) when making an identification from the lineup. During the test phase, participants were shown a perpetrator-present or perpetrator-absent lineup. Participants were then instructed to identify the perpetrator if the perpetrator was present in the lineup or, if the perpetrator was not present in the lineup, choose no one by clicking the "perpetrator is not present" button. Participants then rated their confidence in their identification decision. After the test phase, participants were asked specific details about the mock crime they had witnessed. One of the questions asked specifically about the type of crime committed in the video. Participants who paid attention during the study phase should have answered this question correctly. Because these experiments took place online and outside the confines of a controlled laboratory, correctly answering this question ensured that the participant paid enough attention during the study phase in order to potentially identify the perpetrator from a lineup. Failing to answer this question correctly provided reasonable grounds for exclusion from further analyses.

Calculating Correct ID, False ID, and Filler ID rates.

Correct ID Rate

Because participants witness a crime that contains one perpetrator, each participant makes only one attempt to identify the perpetrator from a perpetrator-present or perpetrator-absent lineup. Thus, a participant can make one correct ID, false ID, or filler ID. The group's correct ID rate is calculated by adding the total number of correct IDs across the entire group of participants and dividing that number by the total number of perpetrator-present lineups. For instance, if 100 participants were presented a perpetrator-present lineup and only 60 participants correctly identified the perpetrator, the correct ID rate would be .60 (i.e. 60/100 = .60).

False ID Rate

Several methods can be used to calculate the false ID rate. If the perpetrator-absent lineup contains a designated innocent suspect, then the false ID rate can be calculated by dividing the total number of false IDs made by the group of participants by the total number of perpetrator-absent lineups presented to that group. For instance, if 100 participants were shown perpetrator-absent lineups and 20 participants falsely identified the innocent suspect, the false ID rate would be .20 (i.e. 20/100 = .20). However, if the perpetrator-absent lineup does not contain a designated innocent suspect, the false ID rate is estimated. One way to estimate the false ID rate is to designate the most often identified filler as the innocent suspect. This method estimates the false ID rate by dividing the number of times this filler is identified by the total number of perpetrator-absent lineups. Another method does not designate an innocent suspect and, instead, estimates the number of false IDs by dividing the total number of filler IDs by the number of fillers presented in a perpetrator-absent lineup. For instance, if participants were shown a perpetrator-absent lineup that consisted of six fillers and participants identified 60 fillers in total, the estimated number of false IDs would be 10 (i.e. 60/6 = 10). In order to estimate the false ID rate, this value is then divided by the total number of perpetrator-absent lineups. If 100 participants were shown these perpetrator-absent lineups, then the estimated false ID rate would be .10. That is, the number of filler IDs (60) is divided by the number of lineup

members (6) which provides the estimated number of false IDs (10). This estimated value is then divided by the total number of perpetrator-absent lineups (100), producing the estimated false ID rate (.10). This is the most often used method to estimate the false ID rate when a designated innocent suspect is not placed in a perpetrator-absent lineup (e.g. Palmer et al., 2013) and this method is used to estimate the false ID rate in this thesis.

Filler ID Rate

The filler ID rate can be calculated separately for perpetrator-present and perpetrator-absent lineups. The filler ID rate for perpetrator-present lineups is calculated by dividing the total number of filler IDs made by the group of participants by the total number of perpetrator-present lineups shown to that group. The filler ID rate for perpetrator-absent lineups can be calculated only if the lineup contains a designated innocent suspect. This is done by dividing the total number of filler IDs from a perpetrator-absent lineup by the total number of perpetrator-absent lineups.

Receiver Operating Characteristic Analysis

Discriminability was measured with receiver operating characteristic (ROC) analysis because differences in discriminability and response bias are evident in ROC curves (Mickes et al., 2012). To construct the ROCs, the correct ID rates for each level of confidence were plotted along the y-axis and the false ID rates for each level of confidence were plotted along the x-axis. This produces a confidence-based ROC curve where each point reflects a correct and false ID rate for a particular level of confidence. A higher ROC indicates greater discriminability because, for any given false ID rate, the higher ROC yields a higher correct ID rate. Standard ROC analysis statistically compares the area underneath the full ROC, which extends across the full range of correct and false ID rates from 0 to 1. In this thesis, however, the false ID rates were estimated in all ROC analyses because there was no designated innocent suspect used in any of the eyewitness identification experiments. Because of this, the range of false ID rates for the confidence-based ROCs extended from 0 to a value less than 1. This means that differences in discriminability were measured by comparing partial area under the curve (*pAUC*) values

for each confidence-based ROC. Unless otherwise stated, the conclusions from ROC analyses are the same using the other methods of calculating false IDs.

Statistically Comparing *pAUC* Values

The *pAUC* values were compared using the statistical package *pROC* (Robin et al., 2011). The *pROC* package is for the statistical computing program R and includes tests for computing and comparing *pAUC* values for two ROC curves. Specificity (1 – false ID rate) was set in the analysis using the smallest, overall false ID rate obtained from either condition. In other words, the false ID rate from the condition that yielded the most conservative responding was used. The bootstrap method was used with the number of replications set to 2,000. This method calculates the *pAUC* values from the ROC curves 2,000 times. These values are then statistically compared using the following formula:

$$D = \frac{(pAUC_1 - pAUC_2)}{s} \quad (1)$$

, where s is the standard deviation of the bootstrap differences and $pAUC_1$ and $pAUC_2$ are the areas under the curve for the two ROCs. Therefore, D is the difference between the two *pAUC* values being compared and is expressed in standard deviation units. In all analyses, alpha was set to .05.

Estimating Lineup Discriminability

ROC analysis can determine whether a difference in discriminability or response bias has occurred (Macmillan & Creelman, 2005), but in many cases ROC analysis cannot be conducted either because the correct and false ID rates for each level of confidence are not collected or there is not enough data to conduct a meaningful ROC analysis. Mickes, Moreland, Clark, and Wixted (2014) have advocated the use of d' in circumstances where ROC analysis cannot be computed. This is a parametric measure of discriminability rooted in signal-detection theory that estimates the *pAUC* value (MacMillan & Creelman, 2005).

If a lineup is "fair" (i.e. a lineup in which the fillers resemble the perpetrator as much as an innocent suspect), then the simplest signal-detection model for lineups consists of two equal-variance Gaussian distributions: a Gaussian distribution for innocent

suspects and a Gaussian distribution for guilty suspects (see Figure 1). Eyewitnesses will have stronger memory for the guilty suspect than the innocent suspect, on average, because the guilty suspect was previously seen. This means that the guilty suspect distribution rests further along the memory strength axis than the innocent suspect distribution. A decision criterion is placed somewhere along the memory strength axis such that memory signals for suspects exceeding that criterion are identified and memory signals for suspects that do not pass that criterion are not identified.

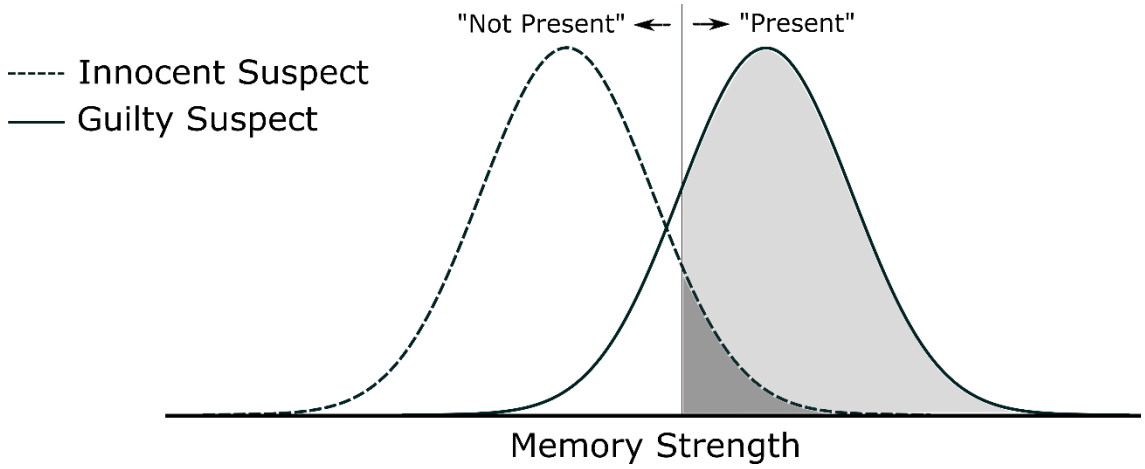

Figure 1. The basic signal-detection model for lineups.

The extent to which the two distributions overlap illustrates an eyewitness's ability to discriminate the guilty suspect from the innocent suspect (Wixted & Mickes, 2014). The amount of overlap can vary depending on the distance between the means of the two distributions (i.e. the separation) and the variance of the two distributions (i.e. the spread). For instance, if the distance between the means of the two distributions is large, then the amount of overlap decreases, illustrating an increase in discriminability. However, if the variance of the two distributions is also large, then the amount of overlap increases, illustrating a reduction in discriminability. Thus, d' is:

$$d' = \frac{\mu_{guilt} - \mu_{inn}}{\sqrt{\frac{1}{2}(\sigma^2{}_{guilt} - \sigma^2{}_{inn})}} \quad (2)$$

, where μ_{guilt} and σ^2_{guilt} is the mean and variance for the guilty suspect distribution and μ_{inn} and σ^2_{inn} is the mean and variance for the innocent suspect distribution, respectively. Because σ^2_{guilt} and σ^2_{inn} are assumed to be equal, the difference in standard deviations between the two distributions no longer needs to be calculated. In the case when the two distributions are assumed to have the same standard deviation, Equation 3 simplifies to:

$$d' = \frac{\mu_{guilt} - \mu_{inn}}{\sigma} \quad (3)$$

, where σ is the standard deviation of the guilty suspect and innocent suspect distributions.

However, it is possible to estimate *d′* using a single correct and false ID rate. For this reason, *d′* was calculated in this thesis using the formula:

$$d' = z(correct\ ID\ rate) - z(false\ ID\ rate) \quad (4)$$

Equation 4 estimates *d′* by converting a single correct and false ID rate into z-scores. A z-score indicates how many standard deviations a value is above or below the mean. Values falling above the mean are transformed into positive z-scores and values falling below the mean are transformed into negative z scores. For values at the mean, the z-score is 0 (because these values do not deviate from the mean). The distribution of correct ID rates (and false ID rates) can range from 0 to 1.00. If we assume that the distribution is Gaussian (as the signal-detection model assumes), then the mean of that distribution is .50. Proportions larger than .50 are converted into positive z-scores and proportions smaller than .50 are converted into negative z-scores. Two proportions that are equally distant from .50 result in the same z-score with alternate signs. For instance, the z-score for a correct ID rate of .75 is .67 and the z-score for a false ID rate of .25 is -.67 (i.e. these values are both equally distant from .50 and, thus, have the same z-score, just with alternate signs). Using these values, *d′* equals 1.35 (i.e. .67 – (-.67) = 1.35). This means that the distance between the mean of the guilty suspect distribution and the mean of the innocent suspect distribution is 1.35 times greater than the standard deviation of the two distributions. Likewise, a *d′* of 2.0 would mean that the distance between the two

distributions (i.e. the separation) is twice as large as the standard deviation of the two distributions (i.e. the spread).

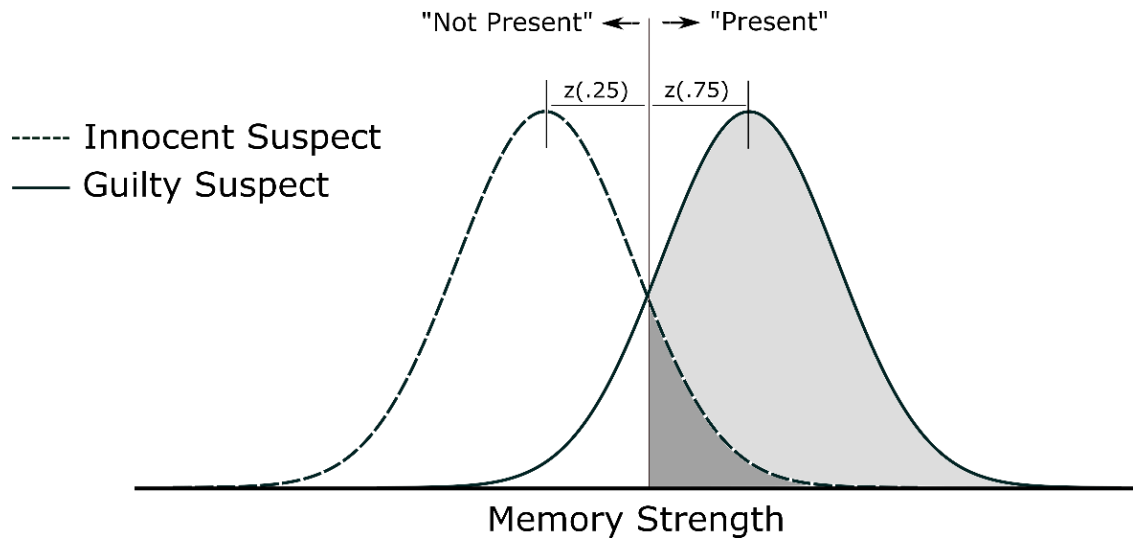

Figure 2. Visualizing *d'* and the signal-detection model. The false ID rate of .25 reflects the dark grey area of the Innocent Suspect distribution. The correct ID rate of .75 reflects the light grey area of the Guilty Suspect distribution. The distance between the means of the innocent suspect and guilty suspect distributions visually depicts discriminability. This distance corresponds to the sum of the z transformed correct ID rate minus the z transformed false ID rate.

To visualize how Equation 4 combines these z-scores into an estimate of discriminability see Figure 2. The light grey area of the guilty suspect distribution reflects a correct ID rate of .75 and the dark grey area of the innocent suspect distribution reflects a false ID rate of .25. We know that transforming these correct and false ID rates into z-scores yield values .67 for the correct ID rate and -.67 for the false ID rate. The z-score .67 represents the distance between the decision criterion and the mean of the guilty suspect distribution. Likewise, the z-score -.67 represents the distance between the decision criterion and the mean of the innocent suspect distribution. Adding these two z-scores, as done in Equation 4, gives the distance between the means of the innocent suspect and guilty suspect distributions, and this distance is a visual depiction of discriminability (see Chapter 1 for review).

Testing for Significance

A G statistic can be used in order to statistically compare two d' values (Gourevitch & Galanter, 1967). For instance, d'_1 and d'_2 are statistically compared using the formula:

$$G = \frac{d'_1 - d'_2}{\sqrt{\mathrm{V}(d'_1)+\mathrm{V}(d'_2)}} \quad (5)$$

, where the numerator is the difference in the d' values and V(d'_1) and V(d'_2) are the variances for both lineups. The denominator is expressed in standard deviation units by taking the square root of V(d'_1) and V(d'_2). Therefore, G is the difference in the d' values being compared and is expressed in standard deviation units. In all analyses alpha was set to .05.

Comparing *d′* and *pAUC*

Mickes et al. (2014) compared d' based statistics based on *G* from Gourevitch and Galanter (1967) with ROC-based statistics based on *D* from Robin et al. (2011). *G* and *D* were calculated using lineup data from several eyewitness identification experiments. The correlation between *G* and *D* was very strong, r = .95. Thus, whether *pAUC* or d' is used to measure eyewitnesses' discriminability from a lineup, the conclusions will often be the same. However, ROC analysis is the preferred method of measuring discriminability because the *pAUC* value is a non-parametric measure of discriminability, unlike d'.

Confidence-Accuracy Characteristic Analysis

In the past, many claims about the weak relationship between confidence and accuracy were made. However, those claims were based on the use of the point bi-serial correlation coefficient, which can mask a strong confidence-accuracy relationship (Juslin et al., 1996). When the data are analysed using calibration analysis, or confidence-accuracy characteristic (CAC) analysis, there is typically a strong confidence-accuracy relationship for individuals who make an identification from a lineup in the laboratory (e.g., Brewer & Wells, 2006; Horry et al., 2012; Sauer et al., 2010; Wixted et al., 2015) and in the field (Behrman & Davey, 2001; Wixted et al., 2015).

CAC analysis most directly supplies the answer to the question that judges and juries have about a testifying eyewitness who has identified a suspect: how accurate is that suspect ID likely to be given the level of confidence that was expressed? CAC analysis plots suspect ID accuracy as a function of confidence using any numerical confidence rating scale. Confidence levels throughout this thesis were binned into low, medium, and high because there were too few responses for certain levels. Suspect ID accuracy for an eyewitness identification is calculated for low, medium, and high confidence levels using the following formula:

$$A = \frac{(\#\ correct\ IDs)}{(\#\ false\ IDs + \#\ correct\ IDs)} \quad (6)$$

. Equation 6 calculates suspect ID accuracy, denoted as *A*, by dividing the number of correct IDs by the total number of suspect IDs. Importantly, filler IDs are excluded from this equation because triers of fact are mainly concerned with the accuracy for suspects. The false IDs were estimated with the same method that was used to conduct ROC analysis.

Suspect ID Accuracy and Prior Probability

The number of correct and false IDs obtained in an experiment is dependent on the prior probability of guilt (i.e. the proportion of lineups that contain the guilty suspect). In this thesis, the guilty suspect base rate was approximately 50% for each eyewitness identification experiment. Suspect ID accuracy reported in these experiments would, in general, be lower if the base rate was lower than 50% and higher if the base rate was higher than 50%. Still, the relationship between two CAC curves would remain so long as the base rate did not vary between conditions. It is important to note that in the real world the guilty suspect base rate is unknown. Yet, there is some indication that suspect ID accuracy for high levels of confidence does not vary greatly across different base rates (Wixted et al. 2016). This means that suspect ID accuracy for at least high levels of confidence calculated in the laboratory (with a base rate of 50%) can serve as a reasonable estimate of suspect ID accuracy in the real world (where the base rate is often unknown).

Computing CAC Standard Errors

The standard errors associated with suspect ID accuracy scores cannot be directly computed and were therefore estimated using a 10,000-trial bootstrap procedure. On each trial, the observed data from perpetrator-present lineups were randomly sampled with replacement to obtain a bootstrap sample of suspect IDs for that trial. For example, if there were 150 high confidence suspect IDs out of 500 lineups, the observed high confidence correct ID rate would equal 150/500 = .30. Thus, on each bootstrap trial, a high confidence suspect ID was registered with probability .30 for each of the 500 lineups (i.e., a high confidence suspect ID would be registered approximately every third lineup, on average). The first bootstrap trial might yield 157 suspect IDs, the next bootstrap trial might yield 141 suspect IDs, and so on. Similarly, on each bootstrap trial, the observed data from perpetrator-absent lineups were randomly sampled with replacement to obtain a bootstrap sample of filler IDs for that trial. For example, if there were 100 high confidence filler IDs out of 500 lineups, the observed high confidence filler ID would equal 100 / 500 = .20. Thus, on each bootstrap trial, a high confidence filler ID would be registered with probability .20 for each of 500 lineups. The first bootstrap trial might yield 94 filler IDs, the next bootstrap trial might yield 101 filler IDs, and so on.

After obtaining a bootstrap sample of suspect IDs and filler IDs on a given bootstrap trial, a suspect ID accuracy score could be computed in exactly the same manner it was computed for the observed data. Thus, for example, if there were 157 suspect IDs and 94 filler IDs on the first bootstrap trial and the size of the lineup was 6, then suspect ID accuracy for the first bootstrap trial would equal 157/(157 + 94/6) = .909. Note that the bootstrap sample of 94 filler IDs was divided by the lineup size to estimate the number of false IDs from perpetrator-absent lineups. Similarly, if there were 141 suspect IDs and 101 filler IDs on the second bootstrap trial, then suspect ID accuracy for the second bootstrap trial would equal 141/(141 + 101/6) = .893. This process was repeated for 10,000 bootstrap trials, and the standard deviation of the 10,000 bootstrap suspect ID scores provided the estimated standard error. The same procedure was followed for each confidence level (i.e. high, medium, and low) in all conditions.

Chapter 3

The US Lineup Outperforms the UK Lineup[1]

The United States and the constituent countries that comprise the United Kingdom (England, Scotland, Northern Ireland, and Wales) have responded similarly to concerns of fallible eyewitness testimony. In the UK, the Devlin Committee chaired by Law Lord Patrick Arthur Devlin investigated a number of criminal cases from the early 1970s in order to draw conclusions on the reliability of suspect IDs. Several years following the Devlin Report (1976), Parliament passed the Police and Criminal Evidence Act (PACE; 1984) which instituted several codes of practice to standardize the methods used to collect eyewitness evidence in England and Wales. Meanwhile, in the US, the Innocence Project has helped release hundreds of individuals who were convicted either solely or in large-part due to false IDs (Innocence Project, 2017). In response to these findings, U.S. Attorney General Janet Reno ordered a panel to discuss and develop national guidelines for collecting eyewitness evidence and, in October 1999, these guidelines were released (Technical Working Group for Eyewitness Evidence, 1999). The guidelines released by the U.S. Department of Justice and the legislation enacted by Parliament (i.e. PACE) advise the police to construct a lineup that consists of one suspect and a number of other fillers (see Chapter 1 for basic introduction to a lineup). Despite sharing these general characteristics, the US lineup and the UK lineup vary quite considerably.

UK Lineup

In 2003, the PACE code of practice was revised in order to favor the use of the Video Identification Parade Electronic Recording (VIPER) database and the Profile Matching (PROMAT) database. These databases have made it possible to create and store large amounts of moving video clips of fillers and suspects to be used in the construction

[1] The experiments in this chapter have been published as Seale-Carlisle, T.M. & Mickes, L. (2016). US line-ups outperform UK line-ups. *Royal Society Open Science, 6*: 160300.

of lineups. Eyewitnesses are now required to view video clips of nine lineup members, one of whom is the police suspect, in a sequence that shows one video at a time. Each video clip starts with the lineup member facing straight towards the camera. The lineup member then rotates their head to the right, displaying the left side of their face, and then rotates their head back towards camera and continues turning their head until the right side of their face is in view. The video clip ends when the lineup member faces towards the camera again. The eyewitness laps through the lineup twice, viewing each lineup member one at a time. After the second lap, they can choose to see any or all of the lineup members again, or attempt to make a decision (by either picking a lineup member or rejecting the lineup). At this stage, they can also choose to view a matrix of all of the lineup members, which involves showing a static image of everyone in the lineup (akin to the simultaneous lineup procedure described below). Approximately half of the police forces in England and Wales and all of the police forces in Scotland use VIPER. The other half of police forces in England and Wales use PROMAT (Valentine, 2006). Both systems produce similar lineups but access a separate database of fillers.

US Lineup

In the US, the traditional lineup procedure is the simultaneous lineup which displays a photo of the police suspect and several fillers to the eyewitness all at the same time. In laboratory studies, the simultaneous lineup has been found to yield high rates of false IDs (Lindsay & Wells, 1980; Wells, 1984). Lindsay and Wells (1985) showed that presenting images sequentially rather than simultaneously could drastically reduce the false ID rate. Their sequential lineup procedure shows photos of lineup members one at a time and requires the eyewitness to decide whether a lineup member is or is not the perpetrator before seeing the next photo. The procedure typically terminates once an identification has been made. Currently, 32% of police jurisdictions in the US opt to use the sequential lineup procedure in place of the simultaneous lineup procedure (Police Executive Forum, 2013).

Mickes et al. (2012) used receiver operating characteristic (ROC) analysis (see Chapter 1 for review on ROC analysis) in order to determine which lineup procedure

yields higher discriminability. Using ROC analysis, Mickes et al. found the US simultaneous lineup procedure yields higher discriminability than the US sequential lineup procedure. A simultaneous advantage has been found by several independent laboratories (Carlson & Carlson, 2014; Dobolyi & Dodson, 2013; Gronlund et al., 2012) and data from two field studies provide converging evidence (Amendola & Wixted, 2014; Wixted et al., 2016).

US Lineup vs. UK Lineup Predictions

Will the US simultaneous lineup yield greater discriminability than the UK lineup? Because the UK lineup presents individuals in sequence, one may expect the simultaneous lineup to yield greater discriminability. However, despite there being a likely simultaneous advantage, there are several other factors that need to be taken into consideration (see Table 1). For instance, the UK lineup utilizes moving rather than static images and consists of nine lineup members rather than just six. The entire UK lineup is also shown twice to the eyewitness whereas the US lineup is only shown once. Because of the myriad of differences, it is difficult to predict which lineup yields greater discriminability. The possible impact of each of these differences is discussed below.

Table 1

The differences between the UK lineup and the US lineup that may impact discriminability are shown in this table.

UK Lineup	US Lineup
Sequential Presentation	Simultaneous Presentation
Moving Images	Static Images
Multiple Laps (Time Restricted)	One Lap (Free Viewing)
9-Person Lineup	6-Person Lineup

Note: It has recently been shown that the simultaneous presentation of faces yields greater discriminability than the sequential presentation of faces. However, there are several other differences between the two lineups that may impact discriminability.

Moving vs. Static Images

An eyewitness to a crime likely sees the face of the perpetrator from multiple angles and so, when an eyewitness is presented a lineup, a moving image of the suspect may contain information which would otherwise be lost in a static image. This extra information could help the eyewitness correctly identify the perpetrator or help remove suspicion from an innocent suspect. Thus, it is reasonable to expect that moving images improve discriminability.

A series of studies conducted by Cutler and colleagues compared static photo lineups with video lineups. After witnessing a video recording of a mock crime, Cutler, Penrod, and Martens (1987) had participants attempt to identify the perpetrator from an enhanced or unenhanced sequential lineup. The enhanced lineup consisted of a close-up (face and shoulders), three-quarter (face and torso view), and full frontal static image of each lineup member. Brief video clips of each lineup member were then shown after the series of static images. The unenhanced lineup only consisted of the close-up static images. Correct and false ID rates were not originally reported for these conditions and, instead, the proportion of correct responses was computed. Cutler et al. note that the effect of enhanced lineup cues had surprisingly no significant effect on proportion correct. In a later study, Cutler and Fisher (1990) had participants witness a live mock crime and compared how participants performed when presented live lineups, video lineups, and photo lineups. The false ID rate was slightly, but significantly, lower for live and video lineups, while the correct ID rate was the same for all three lineup types.

In review of these studies, Cutler, Berman, Penrod, and Fisher (1994) concluded that the additional information present in moving images produced a “trivial effect on identification accuracy” (p. 179) and that “there is no reason to believe that live lineups, videotaped lineups, or photo arrays produce substantial differences in identification performance” (p. 181). However, Valentine and Davis (2015) reanalyzed Cutler et al.’s (1987) data and found that the enhanced lineups (i.e. the lineups containing static and moving images) resulted in a significantly higher correct ID rate and a significantly lower false ID rate than the unenhanced lineups (i.e. the lineups just containing the static images).

Although Cutler et al. (1994) concluded that moving images had a trivial effect on "identification accuracy", their data seem to provide some indication that moving images improve discriminability, though the size of the effect may be small.

There have since been several other studies that have investigated the effects of moving and static images on discriminability from a lineup (e.g. Kerstohlt, Koster, & van Amelsvoolt, 2004; Valentine, Darling, & Memon, 2007; Havard, Memon, Clifford, & Gabbert, 2010; Humphries, Holliday, & Flowe, 2012). Valentine et al. (2007) constructed two UK lineups, one which utilized moving images, as required by PACE, and one which utilized static images. In this experiment, participants witnessed a live theft and were assigned to a perpetrator-present or a perpetrator-absent lineup. The perpetrator was recorded in a VIPER suite at a police station in order to create the moving image. A police identification officer selected fillers that matched the perpetrator on the relevant criteria dictated by PACE (e.g. age, height, and general position in life). The selection of fillers was carried out exactly as it would for an actual police-constructed lineup. The static images were single-shot frames showing a front view of the lineup member's face. Participants watched the entire VIPER lineups from start to finish twice. They were then able to see any lineup member again, as many times as they wished, until they were able to make an identification decision. Thus, the two procedures followed PACE codes of practice except that static images were used in replace of moving images in one of the procedures. The correct ID rate was found to be significantly higher, and the false ID rate to be significantly lower, for the UK lineup containing the moving images than the UK lineup containing the static images, indicating an improvement in discriminability.

Multiple Laps vs. One Viewing

If an overly cautious eyewitness is only given one opportunity to identify the perpetrator, then a potential correct ID could be easily lost. This is particularly problematic for police jurisdictions that have adopted the sequential lineup procedure as this procedure induces conservative responding, which further reduces the correct ID rate (e.g. Steblay et al., 2001; 2011). To remedy this issue, some US jurisdictions have allowed eyewitnesses to view the entire array of lineup members again and, in some cases, multiple "laps" have

been allowed (Klobuchar, Steblay, & Caligiuri, 2006). Although this may help increase the correct ID rate, repeated exposure to the lineup may also interfere with eyewitnesses' memory for the perpetrator (causing a reduction in discriminability). The field data from Klobuchar et al. show that identifications from subsequent laps resulted in roughly the same rate of suspect IDs but also a large increase in filler IDs. An increase in filler IDs reflects an increase in known errors and does not necessarily reflect worse discriminability for the suspect.

Steblay, Dietrich, Ryan, Raczynski, and James (2011) compared correct and false ID rates from sequential single-lap and double-lap lineups. In Experiment 1, participants witnessed a video of a mock crime and attempted to identify the perpetrator after a brief delay. Participants had the option of viewing a second lap after having seen each lineup member, similar to the protocol in Klobuchar et al. (2006). For participants who opted to view a second lap, the correct ID rate increased by 6%, the filler ID rate increased by 14%, and the false ID rate increased by 23%. In Experiment 2, participants were assigned to either single-lap or double-lap sequential lineups. For those assigned to double-lap lineups, the correct ID rate increased by 9%, the filler ID rate increased by 5%, and the false ID rate increased by 15%. Participants who elected (Experiment 1) or were required (Experiment 2) to view the lineup twice made significantly more identifications indicating a liberal shift in response bias. Yet, the large increases in the false ID rate and the relatively smaller increases in the correct ID rate may also indicate a reduction in discriminability.

To statistically compare discriminability for single-lap and double-lap lineups, I computed d' from the overall correct and false ID rates reported by Steblay et al. (2011) and statistically compared these values using the G statistic (see Chapter 2 for review). If necessary, this approach can be used in place of ROC analysis (Mickes et al., 2014). In Experiment 1, participants that saw the lineup once ($d' = 1.55$) had higher discriminability than those who elected to view the lineup twice ($d' = 1.40$), but the difference was not significant ($G = .30$, $p = .76$). In Experiment 2, the effect of repeated laps was directly tested as participants were randomly assigned to single-lap or double-lap lineups. Those who viewed the lineup twice actually had higher discriminability ($d' = 1.40$) than those

who viewed the lineup once ($d' = 1.33$), but again, the difference was not significant ($G = .18$, $p = 0.86$). Viewing the entire lineup a second time likely introduced more interference for the memory of the perpetrator, but also helped participants compare lineup members. When participants are able to compare lineup members, they can become more aware of the features that are diagnostic of guilt (and those features which are not) and can use this information to aid discrimination on the second lap (Wixted & Mickes, 2014). This benefit may have counteracted the interference caused by viewing the lineup again, resulting in no significant difference in discriminability between single-lap and double-lap lineups.

Nine vs. Six Lineup Members

The fairness of a lineup procedure has been a concern of eyewitness identification researchers for quite some time. Researchers have discussed that the showup procedure, which only presents the police suspect to the eyewitness, may be overly suggestive of guilt and may put innocent suspects at too great a risk of a false ID (e.g. Sobel, 1972; Levine & Tapp, 1973). Sobel (1972) has recommended presenting fillers alongside the suspect so as not to make the suspect appear too distinctive. However, Buckhout (1974) showed that innocent suspects that appear distinctive from the fillers (e.g. if the suspect's photo was taken at a different angle than the fillers) are still at risk of a false ID. Researchers have since made a distinction between the number of fillers in a lineup, referred to as the nominal size of a lineup, and the number of fillers in a lineup that resemble the suspect, referred to as the functional size of a lineup (Wells, Leippe, & Olstrom, 1979). Simply increasing the nominal size of a lineup (i.e. by adding fillers that do not resemble the suspect) may not reduce the false ID rate (Malpass, 1981). An eyewitness attempting to identify the perpetrator can essentially ignore fillers that do not resemble the suspect. However, lineups with a greater functional size have been argued to reduce the false ID rate (Wells & Turtle, 1986). If fillers and suspects both resemble the perpetrator, an eyewitness may incorrectly pick a filler instead of an innocent suspect. Note that a reduction in the correct ID rate may occur for the very same reason. The nominal and functional size of the UK lineup is greater than the US lineup. The UK lineup consists of

eight fillers that resemble the suspect, whereas the US lineup only consists of 5 fillers that resemble the suspect. Do lineups with greater nominal and functional size yield greater discriminability?

Cutler et al. (1987) conducted a study that directly manipulated lineup size. After participants witnessed a video recording of a mock crime, they were assigned to either a 6-person or a 12-person lineup. The proportion of correct responses was lower for 12-person lineups, but only when the perpetrator was partially disguised during the crime (note that the perpetrator was not disguised when present in the lineup). Because correct and false ID rates were not reported in this study, it is difficult to determine based on just proportion correct whether discriminability was also affected. Nosworthy and Lindsay (1990) manipulated nominal (Experiment 1) and functional (Experiment 2) lineup size. In both experiments participants witnessed a staged crime and attempted to identify the perpetrator from a lineup after a brief delay. Table 2 contains the correct ID rate, false ID rate, and d' for each lineup condition in Experiments 1 and 2.

In Experiment 1, nominal size was manipulated by adding 3 or 6 fillers to a 4-person lineup. Participants were, therefore, assigned to either a 4-person, 7-person, or 10-person lineup. The additional fillers did not resemble the suspect. This enabled the nominal size to increase without affecting the functional size of the lineup. The additional fillers drew very few filler IDs (only 2.3% of total IDs) indicating that participants virtually dismissed these lineup members. The false ID rates were similar for each lineup, but the correct ID rates were slightly higher for the 7-person and 10-person lineup. Interestingly, discriminability was higher for the 7-person (d' = 1.63) and 10-person (d' = 1.64) lineup than the 4 person-lineup (d' = 1.26). However, there was no significant difference in discriminability between the 4-person lineup and the 7-person lineup (G = .64, p = .51) or the 4-person lineup and the 10-person lineup (G = .70, p = .48).

In Experiment 2, functional size was manipulated by adding fillers that resembled the suspect. In this experiment, participants attempted to identify the perpetrator from a 4-person, 8-person, 12-person, 16-person, or 20-person lineup. Although some of the additional fillers drew a large amount of filler IDs (e.g. one filler in the 20-person lineup

received roughly 30% of incorrect IDs for that lineup), the functional size of the lineups did not significantly impact the correct and false ID rates (as shown in Table 2). However, the false ID rates were already low in the smaller lineup conditions, especially in the 8-person lineup (where the innocent suspect was never identified). Nosworthy and Lindsay attributed these low false ID rates to choosing an innocent suspect that did not look like the perpetrator. The data were reanalyzed by designating the most often identified filler as the innocent suspect. Yet, when the data were reanalyzed this way, the conclusions were still the same; the functional size of the lineup did not significantly impact the false ID rates. Although the UK lineup has a greater functional size than the US lineup, these results suggest that this difference does not significantly impact the correct ID rate or the false ID rate and, therefore, has no impact on discriminability.

Table 2

The lineup size, correct ID rates, false ID rates, and d' values for Experiments 1 and 2 from Nosworthy and Lindsay (1990).

Lineup Size	Correct ID rate	False ID rate	d'
Experiment 1 (Nominal Size)			
4-Person	.47	.09	1.26
7-Person	.53	.06	1.63
10-Person	.67	.09	1.64
Experiment 2 (Functional Size)			
4-Person	.48	.07	1.42
8-Person	.33	.00	1.61*
12-Person	.48	.04	1.70
16-Person	.33	.04	1.31
20-Person	.37	.00	1.72*

Note: For Experiment 2, no participant falsely identified the innocent suspect in the 8-person and 20-person lineups. A standard correction of the false ID rate of 0 was applied in order to estimate d'. The * shows the estimates.

US vs. UK Predictions Revisited

Although the US lineup and the UK lineup differ in several respects, the extant eyewitness identification literature suggests that only two differences may actually impact discriminability: the lineup format (i.e. simultaneous or sequential) and the presentation format (i.e. moving or static images). The advantages are split between the two procedures. The US lineup receives an advantage because the simultaneous lineup procedure has been repeatedly shown to yield greater discriminability than the sequential lineup procedure (e.g. Mickes et al., 2012). Therefore, the US lineup procedure receives a check. The UK lineup receives the other advantage (and the other check) because lineups that contain moving rather than static images have shown a slight, but appreciable improvement in discriminability (Valentine et al., 2007; Valentine & Davis, 2015). The fact that the UK lineup shows nine lineup members, whereas the US lineup only shows six, is unlikely to cause a difference in discriminability. In addition, the fact that the UK lineup shows the entire lineup twice (rather than once) is also not likely to have any impact on discriminability. Because the advantages are split between the two lineup procedures, a direct comparison using ROC analysis is needed in order to determine whether the US lineup or the UK lineup yields greater discriminability.

Confidence-accuracy Characteristic Analysis

The results of ROC analysis are important for policymakers who are charged with deciding which type of lineup to use (Mickes, 2015). However, once a criminal case reaches a court of law, regardless of the procedure that was used during the investigation, and regardless of whether one procedure is shown to have greater discriminability than the other, judges and jurors need to know if high confidence IDs made by eyewitnesses are reliable. That is, they need to know if eyewitnesses who were highly confident are also likely to be accurate. ROC analysis does not provide that answer, but CAC analysis does (Mickes, 2015). The two main comparisons of interest therefore are: discriminability and reliability of the US lineup and the UK lineup. In comparing these lineup procedures, we ask the following questions: 1) which procedure will yield the best discriminability; 2) are confidence and accuracy related for both procedures?

Experiment 1 and 2

Two experiments were conducted to compare discriminability and reliability of US and UK lineups, and they differed only slightly with regard to the UK lineup condition. In one of the experiments (but not the other), after lapping through the lineup twice, participants in the UK condition had the opportunity to view as many lineup members as often as desired before making their decision (PACE, 1984). Because there were no important differences in the results, the data was combined and presented together (but the frequency counts are presented separately in Table 4). We report the results of both ROC analysis, which evaluates the level of discriminability supported by the US and UK lineup procedures, and CAC analysis, which measures the confidence–accuracy relationship associated with suspect IDs for the two procedures. The Appendix contains the ROC and CAC analysis separately for Experiment 1 and Experiment 2.

Methods

Participants

Participants were undergraduate students from the University of California, San Diego and completed the experiment in exchange for course credit (N =2249; 1551 female, 681 male and 17 did not state; age in years: M = 20.62; SD = 2.80). Participants were randomly assigned to the US lineup or UK lineup condition, and to a perpetrator-present lineup or a perpetrator-absent lineup. We determined that a sample size of 1000 (for both Experiment 1 and 2) would yield sufficient power to detect an effect size as large as the one observed in previous research for simultaneous versus sequential line-ups (Mickes et al., 2012). Data collection continued until the term ended. The UCSD Institutional Review Board approved of these experiments.

Materials

The study stimulus was a 20 s video of a mock crime of theft. The perpetrator, a young white male whose face was shown for 8 seconds, stole money and a tablet from a deserted office. A London Metropolitan Police Officer with specialized training in eyewitness identification procedures selected nine foils from the PROMAT database. As

specified by PACE, filler selections were based on the general similarities of the perpetrator's appearance, age, ethnicity, weight, and position in life. The actor in the video was recorded in the same London Metropolitan Identification Suite in the same manner as the fillers. Note that these same fillers were used in the US lineup condition.

Procedure

In Experiment 1 ($N = 962$), participants in the UK condition watched the video of the crime, took part in a 5-minute distractor task (a game of Tetris), and viewed the UK lineup twice before making an identification. The UK lineup procedure consisted of nine videos displayed one at a time in a sequence. Those in the US condition watched the video of the crime, took part in the distractor task, and viewed six photos simultaneously arranged in a 2 X 3 matrix. The position of the perpetrator in the perpetrator-present lineups was randomly determined for each participant. No fillers were designated as the innocent suspect. Confidence in the identification decision was collected using an 11-point scale ranging from 0 (just guessing) to 100 (absolutely certain). Following their identification decision, participants then answered several questions including a validation question (what crime was committed?) and were finally debriefed.

In Experiment 2 ($N = 1{,}288$), participants watched the video of the crime, took part in the same distractor task for five minutes and were either presented a US or UK lineup. Participants in the UK lineup viewed the entire lineup twice and were allowed to repeatedly view any member from the lineup before making an identification decision. As in Experiment 1, those in the US lineup condition viewed a simultaneous lineup presented in a 2 X 3 matrix.

Results

Descriptive Analysis

Those who did not answer the validation question correctly were excluded from all analyses ($n = 44$). The total number of perpetrator-present and perpetrator-absent trials as well as the number of correct IDs, false IDs, and filler IDs for every level of confidence for the UK and the US conditions across Experiments 1 and 2 are shown in Table 3.

Table 3

Response frequencies for every decision outcome for each level of confidence.

	US Lineup					UK Lineup				
	Perp Present			Perp Absent		Perp Present			Perp Absent	
Confidence	SID	FID	noID	FID	noID	SID	FID	noID	FID	noID
					Experiment 1					
0	0	4	1	5	6	0	3	10	7	7
10	1	0	0	4	0	0	3	3	1	1
20	2	3	4	4	4	1	5	2	5	2
30	7	10	6	12	6	1	17	4	6	2
40	10	9	5	14	8	3	15	3	12	5
50	15	13	7	25	11	3	19	12	18	8
60	10	12	8	28	11	5	19	6	17	8
70	17	12	18	27	18	11	27	12	21	10
80	17	12	12	12	13	5	13	7	20	10
90	12	5	12	5	19	4	5	10	11	6
100	6	0	5	2	14	7	4	8	9	5
					Experiment 2					
0	1	3	1	6	2	3	7	6	8	4
10	1	3	3	4	0	0	2	2	1	1
20	5	6	4	4	1	1	5	2	7	1
30	5	10	6	13	12	2	8	10	21	4
40	13	7	5	27	11	5	13	4	14	1
50	12	26	16	26	19	6	19	12	33	8
60	21	11	14	22	14	9	20	8	31	11
70	28	14	18	38	31	16	26	17	44	13
80	16	10	13	19	29	13	28	15	36	19
90	13	4	11	12	16	8	14	10	22	11
100	8	0	8	6	17	7	7	2	9	13

Note: Perp, perpetrator; ID, identification; SID, suspect IDs; FID, filler IDs.

ROC Analysis

The suspect ID rates for perpetrator-present lineups (i.e. correct ID rates), suspect ID rates for perpetrator-absent lineups (i.e. false ID rates) and filler ID rates for both perpetrator-present and perpetrator-absent lineups are shown in Table 4. Because no innocent suspect was designated, the false ID rates were estimated by dividing the number of filler IDs from a perpetrator-absent lineup by the number of filler IDs presented in a perpetrator-absent lineup (i.e. 6 for the US and 9 for the UK). This is the most common method for estimating the false ID rate. There are other methods of estimating the false ID rate (as discussed in Chapter 2), but the conclusions presented in this chapter remain the same regardless of which method is used. The italicized values were used to construct the ROC curves in Figure 1. The bold italicized values are the overall correct and false ID rates that have been traditionally analysed in an effort to determine lineup superiority. However, because both the correct ID rate and the false ID rate are lower for the UK lineup procedure (and could, therefore, mean a shift in responding, not a difference in discriminability), an ROC analysis provides a clearer picture of the discriminability associated with the two procedures.

Table 4

Suspect IDs, filler IDs and no IDs rates by level of confidence per condition for perpetrator-present and perpetrator-absent lineups.

		US Lineup			UK Lineup		
	Confidence	SID	FID	noID	SID	FID	noID
Perp Present	0	***.39***	.30		***.20***	.50	
	10	*.38*	.29		*.19*	.49	
	20	*.38*	.29		*.19*	.48	
	30	*.37*	.27		*.19*	.46	
	40	*.35*	.24		*.18*	.41	
	50	*.31*	.21	.31	*.17*	.36	.30
	60	*.26*	.14		*.15*	.29	
	70	*.20*	.10		*.13*	.22	
	80	*.13*	.05		*.08*	.13	
	90	*.07*	.02		*.05*	.05	
	100	*.02*	.00		*.03*	.02	
Perp Absent	0	***.09***	.45		***.08***	.62	
	10	*.09*	.44		*.07*	.60	
	20	*.09*	.43		*.07*	.59	
	30	*.08*	.42		*.07*	.57	
	40	*.08*	.38		*.07*	.52	
	50	*.06*	.32	.45	*.06*	.48	.30
	60	*.05*	.25		*.05*	.39	
	70	*.03*	.17		*.04*	.30	
	80	*.02*	.08		*.02*	.19	
	90	*.01*	.04		*.01*	.09	
	100	*.00*	.01		*.00*	.03	

Note: Perp, perpetrator; ID, identification; SID, suspect IDs; FID, filler IDs.

Discriminability was measured by conducting ROC analysis and comparing the partial area under the curve (*pAUC*) values for each lineup procedure, rather than computing the AUC for the full ROC. This is because the range of false ID rates for lineup-based ROCs extends from 0 to a value less than 1. In order to calculate the *pAUC*, the most conservative false ID rate obtained from either lineup was selected, which is .08 from the UK lineup. The *pAUC* values were compared using the statistical package pROC (Robin et al., 2011). ROC curves were constructed for the combined US and UK lineup conditions (Figure 1). The *pAUC* for the US lineup (0.017) was significantly greater than the *pAUC* for the UK lineup (0.010), $D = 2.74$, $p = .006$.

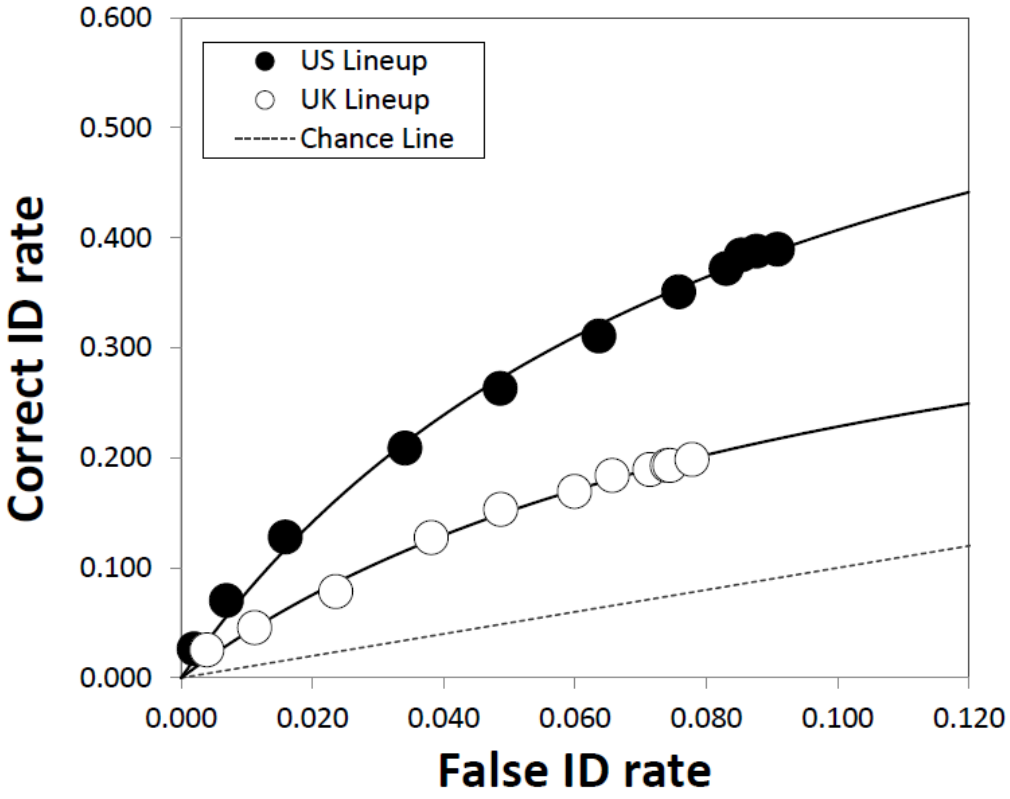

Figure 1. Receiver operating characteristic (ROC) curves for the US lineup and the UK lineup. The dashed line represents chance performance. The lines through the ROC curves were estimated using a hyperbolic function.

Comparing Discriminability of Repeated Viewings

There were some participants who opted to view the lineup members again (n = 128). The majority of those participants only opted to see a lineup member once more (n = 85); very few participants opted to see multiple lineup members again (n = 43), which makes it difficult to estimate discriminability for this group of participants. Thus, we measured whether discriminability for those participants who opted to view lineup members again (n = 128) differed from those who did not. To do so, we computed d' from the overall correct and false ID rates and compared them using the G statistic. We used this approach instead of ROC analysis because separating the data in this manner resulted in too few observations to perform a meaningful *pAUC* analysis (Mickes et al., 2014). Those who viewed lineup members more than the required two times had lower discriminability (d' = 0.35) than those who viewed the lineup members twice (d' = 0.68), but the difference was not significant (G = 1.14, p = 0.253).

CAC Analysis

The relationship between confidence and accuracy was examined using CAC analysis (Mickes, 2015). This analysis only focused on suspect IDs. Filler IDs in perpetrator-present lineups were ignored. This approach provides information most important for judges and jurors as fillers in lineups are already known to be innocent. Examining all of the errors (including filler IDs in perpetrator-present lineups) would, perhaps, be appropriate if testing a psychological theory of calibration. However, for this particular analysis, we sought to determine whether the US and UK lineup procedures differentially affected participants' ability to calibrate confidence and accuracy for the suspect.

The CAC plots are shown in Figure 2. Confidence was binned into low (0-60), medium (70-80), and high (90-100). For each level of confidence, suspect ID accuracy (A) = # correct suspect IDs / (# correct suspect IDs + # innocent suspect IDs). This calculation assumes that the base rates are equal or roughly equal, as they are in this experiment. Because there was no designated innocent suspect, false IDs were divided by the number of fillers present in the perpetrator-absent lineups: nine in the UK lineup and

six in the US lineup. Participants in the UK condition were less able to properly calibrate confidence and accuracy for high, medium, and low confident responses.

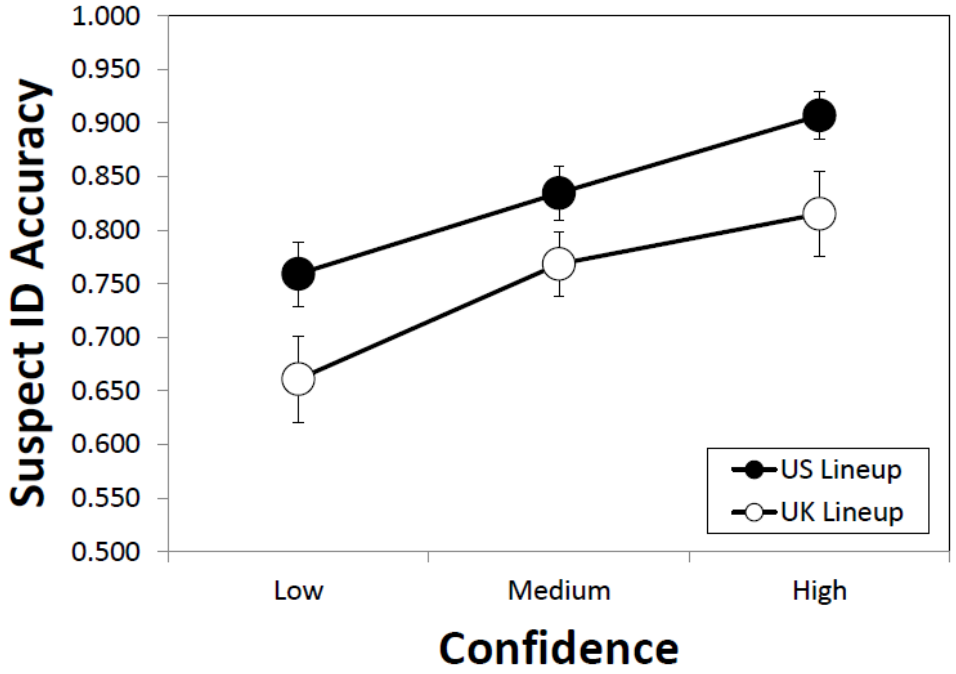

Figure 2. Confidence-accuracy characteristic (CAC) analysis for the US lineup and the UK lineup. The bars represent standard error bars.

General Discussion

Many countries look to the US and UK for adoption of identification procedures. However, the US and the UK identification procedures vary quite considerably. Thus, an important direct comparison between these two identification procedures is needed in order to determine which one is superior. In the first direct comparison of the US and the UK identification procedures using ROC and CAC analyses, the US identification procedure outperformed the UK identification procedure. Specifically, the US identification procedure yielded higher discriminability and significantly higher accuracy at each level of confidence. If these results replicate, many countries including the UK could improve the identification procedure simply by implementing the US simultaneous procedure.

The results from CAC analysis may suggest to some that the UK procedure is superior because eyewitnesses do not underestimate their confidence in their identifications as much as those shown the US procedure. Figure 2 shows suspect ID accuracy across low and medium confidence judgments. Low confidence judgments and medium confidence judgments were made, on average, with 55% and 75% confidence. Suspect ID accuracy for the UK procedure across low and medium levels of confidence provide a close match to these percentages, 65% and 76% respectively. Whereas, those in the US procedure are more "under-confident." Ideally, an eyewitness' confidence matches their suspect ID accuracy, but what is *most* important to the trier of fact is whether the procedure improves suspect ID accuracy even if this results in eyewitnesses becoming under-confident.

It is not possible to pinpoint exactly why the US procedure outperformed the UK procedure because of the array of differences between the two procedures, but these findings could be an example of the often-replicated difference between simultaneous and sequential lineups (e.g. Mickes et al., 2012). Alternatively, the difference in performance between the two procedures may have been caused by the length of time to complete the US procedure compared to the UK procedure. Upon being shown a lineup, participants in the US condition can quickly make an identification decision, whereas participants in the UK condition must wait for the entire lineup to lap through twice before being able to make an identification decision (which takes approximately two minutes). Such a small difference in retention interval is unlikely to account for the drastic differences in discriminability and reliability, but the difference may still be worth considering. Similarly, another possibility is that participants taking part in the experiment online were less engaged during the UK procedure than the US procedure and this may have resulted in worse discriminability and reliability for those in the UK procedure. Lastly, it could be that this pattern of results is due, in part, to the small set of stimuli. Future studies should compare the US and UK lineups using a variety of face stimuli and, if the effect still holds, then each factor should be isolated in order to determine which specific factors affect discriminability and reliability. Given how many innocent and guilty suspects are tested using lineup procedures in both the US and the UK, such work should be an urgent priority.

Chapter 4

The Effect of Descriptions on Discriminability and Reliability

Law enforcement officers often ask an eyewitness to a crime for a detailed description of the perpetrator (Technical Working Group, 1999). An accurate and detailed description is used by law enforcement officers to help apprehend or rule out individuals suspected of a crime and is also used to select appropriate fillers to be placed alongside the suspect in a lineup. A large body of research suggested that a verbal description should contribute to memory performance (Rundus, 1971; Woodward, Bjork, & Jongeward, 1973; Darley & Glass, 1975; Glenberg, Smith, & Green, 1977; Glenberg & Adams, 1978). However, a series of experiments by Schooler and Engstler-Schooler (1990) showed that a verbal description can cause a large reduction in the correct ID rate. This effect is called 'verbal overshadowing' and there have since been numerous attempts to replicate this finding.

Replicating the Verbal Overshadowing Effect

Some researchers have replicated the effect in forensically relevant experiments (e.g. Fallshore & Schooler, 1995; Schooler, Ryan, & Reder, 1996; Dodson, Johnson, & Schooler, 1997; Ryan & Schooler, 1998), but several others have not (e.g. Lovett, Small, & Engstrom, 1992; Yu & Geiselman, 1993; Memon & Bartlett, 2002). Some have observed verbal overshadowing in standard recognition memory tasks (e.g. Brown & Lloyd-Jones, 2002; 2003) yet have also found an effect in the opposite direction (i.e. describing a studied face could facilitate memory; Brown & Lloyd-Jones, 2005). A meta-analysis (Meissner & Brigham, 2001) found a small, statistically significant verbal overshadowing effect, a report less robust than Schooler and Engstler-Schooler's (1990) original results, but the validity of the effect still remained in question. This is because the meta-analysis included studies that used a variety of stimuli, delays, filler tasks, and other materials, and many of the studies included in the meta-analysis either found no effect or an effect in the opposite direction. The first registered replication report (Alogna et al., 2014), a concerted effort of 31 independent laboratories, attempted to directly replicate

the initial findings from Schooler and Engstler-Schooler (1990). The meta-analytic effect across the 31 studies found the size of the effect to be substantially smaller than the initial findings; still, a statistically significant verbal overshadowing effect was observed.

Theoretical Implications of Verbal Overshadowing

Despite successfully replicating the original report (Alogna et al., 2014), because there was no way to measure false ID rates (as the perpetrator was always present in the lineup), the particular effect verbalization has on identification performance remains unclear (Mickes & Wixted, 2015). That is, it is unclear whether the reduction in the correct ID rate was caused by conservative responding (i.e. a change in response bias) or because memory was adversely affected (i.e. a reduction in discriminability). Determining whether verbal overshadowing reflects a change in response bias, discriminability, or both has important theoretical implications (Mickes & Wixted, 2015; Rotello et al., 2014). There are three main theories of the verbal overshadowing effect. First, the content account (e.g., Meissner et al., 2001) holds that the verbal description interferes with the memory of the perpetrator, causing a reduction in discriminability. Second, the criterion-shift account (Clare & Lewandowsky, 2004) holds that verbal overshadowing reflects a change in response bias rather than a change in discriminability. Lastly, the processing account holds that the switch from visual to verbal processing (Schooler, 2002) affects both discriminability *and* response bias (Chin & Schooler, 2008). Each theory is discussed below.

Content Accounts of Verbal Overshadowing

Recoding Interference

Schooler and Engstler-Schooler (1990) argued that verbal overshadowing is caused by recoding interference. According to this theory, participants that attempt to describe a previously seen stimulus "recode" the visual memory of the stimulus into a verbal memory. However, the verbal memory may become impoverished when participants fail to describe key components of the stimulus. During a recognition task,

the impoverished verbal memory interferes with the visual memory causing a reduction in discriminability.

In a series of experiments, Schooler and Engstler-Schooler showed that recoding interference occurs when participants describe a previously seen face. In Experiments 1 and 2, participants watched a video of a mock bank robbery and were either instructed to describe the bank robber in as much detail as possible or complete a control task before attempting to identify the bank robber from a perpetrator-present lineup. In both experiments, participants who had described the perpetrator were less likely to correctly identify the perpetrator from the lineup compared to control participants. Schooler and Engstler-Schooler hypothesized that the impoverished verbal memory of the face interfered with the rich visual memory of the face during the recognition task. The verbal memory was impoverished because it was difficult to adequately describe the perpetrator's face. Participants relying on this impoverished verbal memory were less able to identify the perpetrator as a result.

Schooler and Engstler-Schooler found that recoding interference applies to other hard-to-describe stimuli as well. In Experiment 3, participants who studied and attempted to describe a particular colour were less likely to recognize that same colour compared to participants who did not provide a description. However, when participants described an easy-to-describe stimulus, recoding interference was reduced. In Experiment 4, participants watched the video of the bank robbery. After the video, participants were instructed to either describe the perpetrator's face, complete a control task, or recall and write down the perpetrator's spoken statement (at the beginning of the video, the perpetrator says to the bank clerk, "Just follow the instructions, don't press the alarm, and you won't get hurt"). Each group of participants then attempted to identify the perpetrator from the lineup and attempted to identify the statement from the video. The participants who described the perpetrator's face experienced verbal overshadowing, replicating previous findings from Experiments 1 and 2. Those who described the spoken statement, however, did not experience verbal overshadowing. In fact, these participants had better memory for the spoken statement relative to controls. These results showed that verbal

overshadowing is particularly associated with memories that cannot be easily put into words. When the studied stimulus is easy to describe, memory may actually improve.

Whether a stimulus is easy or hard to describe depends in part on 1) the ability of the participant to perceive the stimulus and 2) the ability of the participant to translate the various nuances of those perceptions into words. Recoding interference predicts a large verbal overshadowing effect for "individuals who possess a level of perceptual expertise that exceeds their verbal ability" (p. 399, Chin & Schooler, 2008). These individuals will be ill-equipped to effectively translate their rich perceptions of a stimulus into a rich verbal memory of the stimulus. Melcher and Schooler (1996) tested the verbal overshadowing effect across three groups of participants that had varying levels of perceptual and verbal expertise. They had participants taste and attempt to describe a glass of wine prior to a recognition task where the same wine and several other new wines were presented. Participants were either trained wine drinkers (professionals or had taken a wine seminar), untrained wine drinkers, or non-wine drinkers. Verbal overshadowing was only observed for the untrained wine drinkers, consistent with the recoding interference predictions. Those trained in wine drinking had the verbal ability to effectively translate those perceptions into words, whereas non-wine drinkers neither had much ability to perceive the taste of the wine nor describe the taste of the wine. However, the perceptual ability of untrained wine drinkers greatly exceeded their ability to describe the wine and they were less able to recognize the wine as a result.

Retrieval-based Interference

Retrieval-based interference theory holds that the inaccurate details within the description interfere with the memory of the perpetrator, causing a reduction in discriminability (Meissner, Brigham, & Kelley, 2001). During a lineup identification task, for example, retrieval-based interference occurs when participants retrieve inaccurate details from a previous description and mistakenly rely on those details to make an identification. If participants provide a large number of inaccurate details within their description, the amount of interference is predicted to increase and cause a greater reduction in discriminability. In a series of experiments, Finger and Pezdek (1999) had

participants witness a mock crime and later attempt to identify the perpetrator from a perpetrator-present lineup. Participants were given either elaborative description instructions based on the Cognitive Interview (Geiselman et al., 1984) or standard police description instructions. The Cognitive Interview is a method of interviewing eyewitnesses using several retrieval strategies that encourage elaborate and extensive recall of crime-related details; participants are encouraged to provide crime-related details even if they are somewhat unsure in their accuracy for those details (Geiselman et al., 1984). Finger and Pezdek found that participants in the elaborative description condition (i.e. the Cognitive Interview condition) provided more correct *and* incorrect details about the perpetrator and were less able to correctly identify the perpetrator from a lineup (i.e. evidence of verbal overshadowing).

If increasing the number of inaccurate details within the description increases the amount of retrieval-based interference, then limiting the number of inaccurate details within the description should decrease the amount of interference. In a study by Meissner et al. (2001), participants witnessed a mock crime and were either encouraged to describe the perpetrator in as much detail as they could provide (i.e. elaborative description instructions) or were instructed to only provide details of the face they could accurately remember. By mostly providing accurate details of the perpetrator, participants should provide fewer inaccurate details of the perpetrator and, consequently, reduce the amount of retrieval-based interference. Consistent with this prediction, those given elaborative description instructions provided many more correct and incorrect details, and were less able to correctly identify the perpetrator from a lineup. Whereas, those who were instructed to strive for accuracy rather than quantity of crime-related details provided fewer correct and incorrect details, and were better able to correctly identify the perpetrator.

Content Account Limitations

The content accounts of verbal overshadowing predict a strong correlation between description accuracy and discriminability. Consider an eyewitness that has described the perpetrator's face and is attempting to identify the perpetrator from a lineup.

There should be less recoding interference when the eyewitness accurately describes the perpetrator. An eyewitness who accurately describes the perpetrator should "recode" the visual memory of the perpetrator into a reliable verbal memory which would be useful in discriminating the perpetrator from the innocent suspect. Similarly, there should be less retrieval-based interference, if the description of the perpetrator is highly accurate. This is because there will be few (if any) inaccurate details that the eyewitness could retrieve and use to make an identification. However, contrary to these predictions, the correlation between description accuracy and discriminability has not often been found. Although a few have found such a correlation (Meissner, Sporer, & Susa, 2008), the majority have not (e.g. Schooler & Engstler-Schooler, 1990; Fallshore & Schooler, 1995; Kitigami, Sato, & Yoshikawa, 2002; Brown & Lloyd-Jones, 2003). It is difficult to ascribe the verbal overshadowing effect to the quality of the verbal descriptions when there is often no observed relationship between the two.

Recoding interference and retrieval-based interference predict reduced discriminability for the verbalized stimulus; discriminability should remain unimpaired for other studied stimuli (as these stimuli were not described). Yet, verbal overshadowing has, in fact, been observed for stimuli that were not previously described. Dodson et al. (1997) had participants study a male and a female face. Afterwards, participants either described the male face, described the female face, or engaged in a control task. Participants who described only the male face or described only the female face were less likely to correctly identify *both* faces compared to control participants. Similarly, Westerman and Larsen (1997) had participants study photographs of a car and a face, but were instructed to either describe the car or engage in a control task. Participants who described the *car* were less likely to correctly identify the *face* compared to controls. It is difficult for the content accounts to explain why describing a previously seen stimulus such as a car impacts discriminability for an unrelated item such as a face.

Transfer-Inappropriate Processing Account

To accommodate these findings, Schooler, Fiore, and Brandimonte (1997) proposed an alternative account of verbal overshadowing, one which stems from the

transfer-appropriate processing and retrieval-induced forgetting frameworks. According to this account, verbally describing a previously seen stimulus causes a general, but temporary, shift in how participants process information. Schooler et al. regard this as a transfer-*inappropriate* processing shift that ultimately causes a reduction in discriminability for described and non-described stimuli, regardless of whether the description is highly accurate. The transfer-appropriate processing framework and the retrieval-induced forgetting framework are briefly discussed before a more detailed discussion of the transfer-*inappropriate* processing account.

Transfer-appropriate Processing Framework

The premise of the transfer-appropriate processing framework is that performance on a memory test benefits most when the processes used to encode the stimuli overlap maximally with the processes used to retrieve the stimuli (e.g. Roediger, Weldon, Challis, 1989; Roediger, 1990). Processing mismatches between these two stages can result in retrieval failure. Morris, Bransford, and Frank (1977) demonstrated this by manipulating the processes used during encoding and retrieval for a list of words. During encoding, participants were presented with sentences designed to encourage semantic or phonemic encoding of a particular word. For example, participants were either presented "Eagle is a large bird" or "Eagle rhymes with legal" and were instructed to answer *yes* or *no* to each statement. During retrieval, participants took part in a standard recognition test where old words (e.g. eagle) and new words were randomly intermixed and presented one at a time to the participant for an *old* or *new* response. Participants then took part in a rhyme recognition test, which required participants to discriminate words that rhymed with old words from words that did not rhyme with old words (e.g. regal would be considered *old* as it rhymes with eagle). Results from the standard recognition test showed that participants who encoded the semantics of the word (e.g. eagle is a large bird) had better memory than participants who encoded the phonemic qualities of the word (e.g. eagle rhymes with legal). However, results from the rhyme recognition test showed the opposite pattern: those who encoded the phonemic qualities of the word had better memory in this test than those who had encoded the semantics of the word. The results by Morris et al.

highlight the importance of matching the processes used during the encoding and retrieval stages. The demands of the encoding and retrieval tasks can impact memory-related processes in a way that may hinder or facilitate memory performance. It is important to consider whether the task of verbally describing a previously seen face, for example, encourages retrieval processes that match well with the visual processes used to encode the face.

Retrieval-induced Forgetting Framework

Schooler et al. (1997) combined the transfer-appropriate processing framework with the idea of retrieval-induced forgetting. It has been shown that the act of retrieving information from memory can inhibit access to non-retrieved information, essentially causing forgetting of non-retrieved information. Anderson, Bjork, and Bjork (1994) demonstrated this by having participants study a list of category-exemplar pairs. For example, participants studied fruit-orange, fruit-apple, tree-hickory, and tree-maple. Following the study phase, participants practiced retrieving some of these items from a subset of categories, but did not practice retrieving other items from other categories. For instance, participants practiced retrieving fruits (i.e. by filling in fruit-or___ with "orange"), but did not practice retrieving items related to trees. Because fruits received retrieval practice they were labeled as *RP* (i.e. retrieval practice category), whereas trees were labeled as *NRP* (i.e. non-retrieval practice category) because they did not receive retrieval practice. Although some fruits received retrieval practice, some other fruits, such as "apple", were not practiced. These items were labeled *RP-* because they belonged to a practiced category but did not receive retrieval practice. After a distractor task, participants were then asked to recall as many exemplars from the original study list as they could remember. The *RP* items (e.g. orange) were recalled more often than the *NRP* items (e.g. hickory), as one might expect. What is surprising is that the *NRP* items were recalled more often than the *RP-* items (e.g. apple). This result shows that retrieving a subset of items belonging to a specific category can inhibit recall of the other items belonging to that same category. In other words, retrieving "orange" from the study list inhibited the retrieval of

"apple". Had participants not practiced "orange", then "apple" would have been recalled as often as "hickory" or "maple" were recalled.

Evidence for Transfer-inappropriate Processing

Combining these two frameworks, Schooler et al. (1997) proposed that verbal overshadowing is caused by transfer-inappropriate processing. This account is based on the following four premises: 1) instructing participants to describe a previously seen stimulus encourages the retrieval of the verbalizable aspects of the stimulus. This premise follows from the transfer-appropriate processing framework which holds that retrieval processes are determined, in part, by the demands of the task (e.g. Morris et al., 1977; Roediger et al., 1989). Retrieving the verbalizable aspects of the stimulus 2) inhibits the retrieval of the non-verbalizable aspects of the stimulus. This premise is supported by research showing that retrieval of information belonging to a particular category inhibits the retrieval of other information belonging to that same category (e.g. Anderson et al. 1994; Anderson & Spellman, 1995). This interference is not isolated to the described stimulus, but 3) is broad in scope and impairs non-verbal processing of other non-described stimuli. Together, these three premises explain why discriminability is worse for studied items that were not described (e.g. Westerman & Larson, 1997; Dodson et al., 1997) and why discriminability can be worse for studied items even when the description is highly accurate (e.g. Schooler & Engstler-Schooler, 1990; Schooler & Fallshore, 1995). That is, according to this account, verbalizing a previously seen stimulus causes a general shift in processing which broadly impacts discriminability for both described and non-described studied items. The reduction in discriminability can be reversed 4) if participants engage in a task that encourages "appropriate" processing – a task that encourages the retrieval of the non-verbalizable aspects of a stimulus.

Finger (2002) tested whether verbal overshadowing can be reversed by engaging in an unrelated perceptual task prior to taking a recognition test. In Experiment 1, participants watched a video of a mock crime. After the video, participants either described the perpetrator or took part in a control task. Participants then engaged in either a non-verbal processing task in which they completed a series of mazes or a verbal

processing task in the form of a questionnaire. The maze task required participants to draw a line from the start of the maze to the correct end point. Those given the questionnaire were instructed to recall as many category exemplars as possible from six large categories. Participants then attempted to identify the perpetrator from a six-person lineup. In Experiment 2, the procedure was the same procedure used in Experiment 1, but the non-verbal processing task was different. The maze task was replaced with a music task in which participants listened to an instrumental piece of music and had to count the number of times a particular tone was presented. Participants who had engaged in the maze task (Experiment 1) and the music task (Experiment 2) correctly identified the perpetrator more often than those who had took part in the questionnaire. In fact, these participants performed as well as those who had not previously described the perpetrator. These results suggest that verbal overshadowing occurs as a result of widespread disruption of perceptual processes (i.e. a shift towards verbal processing). This disruption can be corrected by engaging in a perceptual task, such as a maze or a music task, prior to taking a recognition test (i.e. shifting participants back towards non-verbal processing).

Finger (2002) showed that it is possible to reverse verbal overshadowing by engaging in an unrelated perceptual task, but can an unrelated perceptual task also contribute to verbal overshadowing? If describing the perpetrator encourages the use of verbal processes which are not conducive to identification, perhaps there are other tasks that can encourage the use of verbal processes as well. Macrae and Lewis (2002) had participants study a video of a mock crime. Following the video, participants *did not* describe the perpetrator, but rather engaged in either a global or local processing task. Face recognition has been shown to rely heavily on global processes (e.g. Tanaka & Farah, 1993). Engaging in a task that encourages the use of global processing might facilitate face recognition, whereas engaging in a task that discourages the use of global processing (i.e. a local processing task) might impair face recognition (i.e. might lead to a verbal overshadowing effect). Participants were presented with a long list of Navon (1977) letters. These are large letters (e.g. T) that are comprised of smaller different letters (e.g. y). Participants either had to report the large letter (global processing task) or the small letter (local processing task) or read an unrelated text for the same duration (control task).

Participants then attempted to identify the perpetrator from a perpetrator-present lineup. Macrae and Lewis found that participants who shifted towards local processing (i.e. who reported the smaller letters) were less likely to correctly identify the perpetrator than control participants. In effect, verbal overshadowing occurred even though participants had not described the perpetrator. Yet, those who shifted towards global processing (i.e. who reported the large letters) were better at correctly identifying the perpetrator than control participants. Results from Finger (2002) and Macrae and Lewis (2002) demonstrate that the specific processing operations utilized in between study and test can either cause verbal overshadowing or reverse verbal overshadowing.

Criterion Shift Account

Although eyewitnesses are required to describe the perpetrator to law enforcement officers (Technical Working Group, 1999), the task of verbally describing the perpetrator is, nevertheless, challenging. Clare and Lewandowsky (2004) argued that participants who find this task especially difficult may believe that their memory for the perpetrator is poor and, as a result, may be more cautious when attempting to make an identification (i.e. may be more conservative in making an identification). This theory makes two key predictions: participants who had described the perpetrator will be less likely to correctly identify the perpetrator from the lineup (i.e. fewer correct IDs) *and*, when the perpetrator is absent from the lineup, participants will be less likely to choose an innocent suspect (i.e. fewer false IDs). Most of the verbal overshadowing literature *has not* provided perpetrator-absent lineups (e.g. Schooler & Engstler-Schooler, 1990; Westerman & Larsen, 1997; Finger & Pezdek, 1999; Macrae & Lewis, 2001) – making it impossible to measure false IDs. Whereas, the few studies that have shown perpetrator-absent lineups have not provided participants the option to reject the lineup (e.g. Dodson et al., 1997; Ryan & Schooler, 1998) which makes these false IDs useless in determining whether a change in response bias has occurred.

In Experiment 1, Clare and Lewandowsky (2004) had participants watch a video of a mock crime and randomly assigned participants to one of three conditions: participants either described the specific features of the perpetrator, described the holistic

features of the perpetrator, or completed a control task. The specific feature task required participants to fill out a questionnaire regarding the perpetrator's hair colour, eye colour, nose, mouth, ears, and so on. Participants assigned to the holistic feature task were given a questionnaire asking participants to rate the perpetrator's intelligence, friendliness, honesty, and the "averageness" of the perpetrator's face. Those in the control condition had to list exemplars of several unrelated categories. Participants then attempted to identify the perpetrator from either a perpetrator-present or perpetrator-absent lineup. For each lineup, participants had the option to either identify a lineup member or reject the lineup in the event the perpetrator was absent. By providing these options to participants, Clare and Lewandowsky could properly measure the correct and false ID rates.

They predicted that participants in the featural and holistic conditions would make fewer correct *and* false IDs than control participants. The correct ID rate for participants in the featural (.69) and holistic conditions (.57) was lower compared to control participants (.80). That is, a verbal overshadowing effect was observed for those who had described the perpetrator's face. However, the question of interest was whether this reduction in the correct ID rate was also accompanied by a reduction in the false ID rate. If so, this pattern would indicate that these participants were less likely to make an identification (i.e. a conservative shift in decision criterion). If, however, the false ID rate increased while the correct ID rate decreased, then a reduction in discriminability has occurred. This is what the content and transfer-inappropriate processing accounts of verbal overshadowing predict (i.e. because these participants described the perpetrator, discriminability should be worse). The false ID rate for those in the featural condition (0.00) was lower compared to control participants (.05), indicating a conservative shift in decision criterion. Yet, the false ID rate for those who had described the holistic features (.20) was higher compared to control participants (.05), indicating a reduction in discriminability.

Criterion Shift Account Limitations

Altogether, these findings provide mixed support for the criterion shift account. Although a conservative shift in decision criteria occurred for those in the featural

condition, it is not clear why a conservative shift did not also occur for those in the holistic condition. In both of these conditions participants described the perpetrator, and, according to Clare and Lewandowsky, both groups of participants should be more conservative in making an identification. Perhaps participants only experienced difficulty when attempting to describe the specific features of the perpetrator's face and did not experience difficulty when attempting to describe the holistic features of the perpetrator. Participants in the holistic condition simply had to rate how friendly, honest, attractive, etc. the perpetrator looked. This may have been too easy of a task. Participants might have experienced difficulty if they actually had to explain why the perpetrator looked friendly, for example. Such a task might have caused participants to become more conservative when making an identification.

To statistically compare discriminability for control, featural and holistic conditions, d' was computed from the overall correct and false ID rates reported by Clare and Lewandowsky and statistically compared using the G statistic (see Chapter 2 for review). Discriminability was significantly worse for those in the holistic condition ($d' = 1.03$) compared to those in the control condition ($d' = 2.53$, $G = 2.49$, $p = .01$), but there was no significant difference in discriminability for those in the featural condition ($d' = 2.83$) compared to those in the control condition ($d' = 2.53$, $G = .31$, $p = .75$). It is not clear why discriminability was only reduced in the holistic condition and not also reduced in the featural condition. Both the transfer-inappropriate processing account and content accounts of verbal overshadowing predict a reduction in discriminability for both conditions. As it currently stands, it remains unclear whether verbal overshadowing is caused by a shift in response bias or a reduction in discriminability. Further analysis is needed that specifically measures response bias and discriminability.

Experiment 3

To determine whether verbal overshadowing reflects a change in response bias, discriminability, or both, a replication of Experiment 1 from Schooler and Engstler-Schooler (1990) was conducted. Importantly, a perpetrator-absent lineup condition was included in order to measure false ID rates. With both correct and false ID rates, receiver

operating characteristic (ROC) analysis can then be conducted (see Chapters 1 and 2 for review; Wixted & Mickes, 2012; Gronlund et al., 2014; National Research Council, 2014).

Reliability of an Identification

If verbal overshadowing reduces discriminability (producing a lower ROC) and not just a conservative shift in responding, Alogna et al. (2014) wrote that suspect IDs admitted as evidence in court "…should be weighted less if the witness had provided a description earlier" (p. 557). Mickes (2015) made the point that because ROC analysis does not measure the reliability of a suspect ID (i.e. ROC analysis is not a measure of probative value), a more appropriate analysis for judges and jurors is confidence-accuracy characteristic (CAC) analysis. It can answer the question: is an identification made with high confidence likely to be accurate? In many cases, people can appreciate conditions that affect their memory and adjust their confidence accordingly (e.g., if they only saw the perpetrator for a short duration, they tend to be less likely to give high confidence identifications, but those high confidence identifications tend to be highly accurate; e.g. Palmer et al., 2013). An eyewitness may or may not appreciate the fact that providing a verbal description can impact memory accuracy. CAC analysis was conducted in order to test whether confidence and accuracy are related across all levels of confidence.

Methods

Participants

Undergraduate students (N = 780) at the University of California, San Diego (UCSD) participated online for course credit. According to results of a power analysis, based on the ROC results reported in Experiment 1 of Mickes et al. (2012), a total sample size of 780 was estimated to detect a difference between the groups with 80% power. Participants (n = 63) reported that they previously viewed the video and were therefore not included in the analyses. Of the remaining (n = 717; 472 female, 239 male, and 6 did not specify; age in years: M = 20.5, SD = 2.55). Participants were randomly assigned to the control condition or the verbal condition, and were tested on a perpetrator-present ($n_{control}$ = 171; n_{verbal} = 190) or a perpetrator-absent lineup ($n_{control}$ =

188; n_{verbal} = 168) based on random assignment. The UCSD Institutional Review Board approved of this experiment.

Materials

Many of the materials used throughout the study, including the videotape of the bank robbery, the instructions throughout the experiment, and the photographs of the lineup members, were the same materials used in the original (Schooler & Engstler-Schooler, 1990) and replicated reports (Alogna et al., 2014). Schooler provided a digitized version of these materials to incorporate for computer use. The original crossword puzzle given to participants during the distractor task in the original report was no longer available, so Schooler selected a comparable crossword puzzle for the replicated report, which was also used in this experiment.

Procedure

As shown in Figure 1, participants studied a video of the bank robbery and, immediately following the video, either provided a verbal description of the perpetrator (i.e. the bank robber) or completed a control task (i.e. listed countries and their corresponding capitals). Participants then engaged in a 20-minute distractor task (i.e. attempted to solve the crossword puzzle) before viewing a lineup. The lineup either did or did not contain the target. The size of the lineup was modified from eight to six members in order to randomly assign participants to perpetrator-absent and perpetrator-present lineup conditions. This was done so that the eight original photos of the lineup members could still be used. If the lineup consisted of eight members, then a new filler would be needed to replace the perpetrator in the perpetrator-absent lineup. Confidence in the identification decision was also collected in order to conduct ROC and CAC analyses. It is important to note that in the original study (Experiment 4; Schooler & Engstler-Schooler, 1990) a between-subjects condition was included, but was excluded in this study. Also, the delay (i.e. length of the distractor task) was 10 minutes in the original study as opposed to 20 minutes in this study. These changes were made in order to follow the exact procedure used in Alogna et al. (2014).

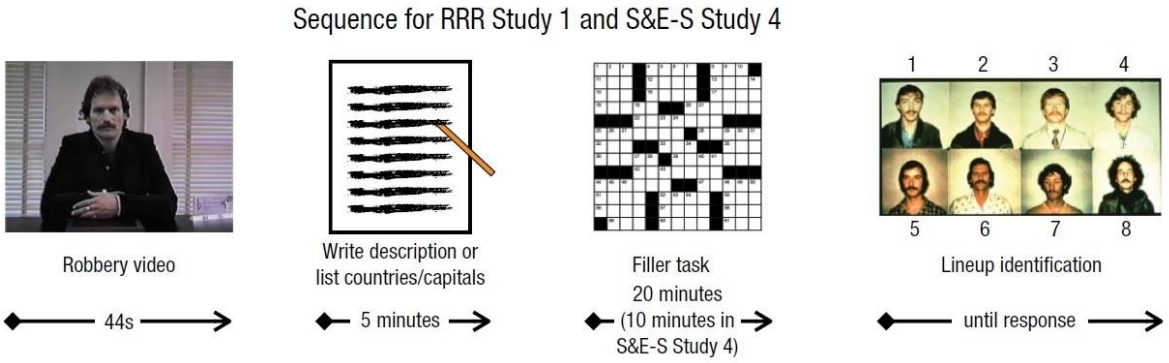

Figure 1. Illustration of the task sequence for the first registered replication report (RRR Study 1) from Alogna et al. (2014).

Results

Descriptive Analysis

The total number of perpetrator-present and perpetrator-absent trials as well as the number of correct IDs, false IDs, and filler IDs for every level of confidence for the verbal and the control condition are shown in Table 1.

Table 1

Response frequencies displayed for every decision outcome for each level of confidence.

	Verbal Condition					Control Condition				
	Perp Present			Perp Absent		Perp Present			Perp Absent	
Confidence	CID	FID	noID	FID	noID	CID	FID	noID	FID	noID
1	1	0	2	2	1	3	2	1	5	2
2	3	7	2	5	6	1	4	1	11	4
3	5	10	3	13	3	10	6	8	22	6
4	18	12	14	20	13	17	15	12	35	14
5	24	14	12	32	30	19	13	7	35	20
6	31	5	5	14	17	21	9	3	14	8
7	16	3	3	5	7	13	4	2	6	6

Note: Perp, perpetrator; ID, identification; CID, correct IDs; FID, filler IDs.

ROC Analysis

The suspect ID rates for perpetrator-present lineups (i.e. correct ID rates), suspect ID rates for perpetrator-absent lineups (i.e. false ID rates) and filler ID rates for both perpetrator-present and perpetrator-absent lineups for each level of confidence are shown in Table 2. The original (Schooler & Engstler-Schooler, 1990) and replicated reports (Alogna et al., 2014) observed a 22% [95% confidence interval: -44% to -0.01%] and 4% [95% confidence interval: -7% to -1%] reduction in correct ID rate due to verbalization, respectively. We observed an 8% [95% confidence interval: -4% to 20%] increase in correct ID rate, but the confidence interval for the correct ID rate findings, ignoring false ID rates, overlap with the confidence intervals for both the original and replicated reports.

Table 2

Identification rates for perpetrator-present and perpetrator-absent lineups.

		Perp Present			Perp Absent		
	Confidence	Correct ID	Filler ID	no ID	False ID	Filler ID	no ID
Verbal	1	***.52***	.27		***.09***	.54	
	2	*.51*	.27		*.09*	.53	
	3	*.50*	.23		*.08*	.50	
	4	*.47*	.18	.22	*.07*	.42	.46
	5	*.37*	.12		*.05*	.30	
	6	*.25*	.04		*.02*	.11	
	7	*.08*	.02		*.01*	.03	
Control	1	***.49***	.31		***.11***	.68	
	2	*.47*	.30		*.11*	.65	
	3	*.47*	.28		*.10*	.60	
	4	*.41*	.24	.20	*.08*	.48	.32
	5	*.31*	.15		*.05*	.29	
	6	*.20*	.08		*.02*	.11	
	7	*.08*	.02		*.01*	.03	

Note: Perp, perpetrator; ID, identification.

To measure whether verbalization affects discriminability, and not just response bias, ROC analysis was conducted. The italicized values in Table 2 were used to construct the ROC curves shown in Figure 2. Because lineup-based ROCs extend from 0 to a value less than 1, *pAUC* was calculated (Mickes, et al. 2012) by using the false ID cutoff of .082 (i.e. 1 - .082 = .918) for both conditions. The *pAUC* for the control condition (0.036) was not significantly different than the *pAUC* for the verbal condition (0.027), D = -0.45, p = .65. Thus, the ROC curves reveal that discriminability in the verbal condition was not significantly different than discriminability in the control condition. Note that the Appendix contains the ROC analysis using the other method of estimating the false ID rate (see Chapter 2). The conclusions are the same regardless of which method is used to estimate the false ID rate.

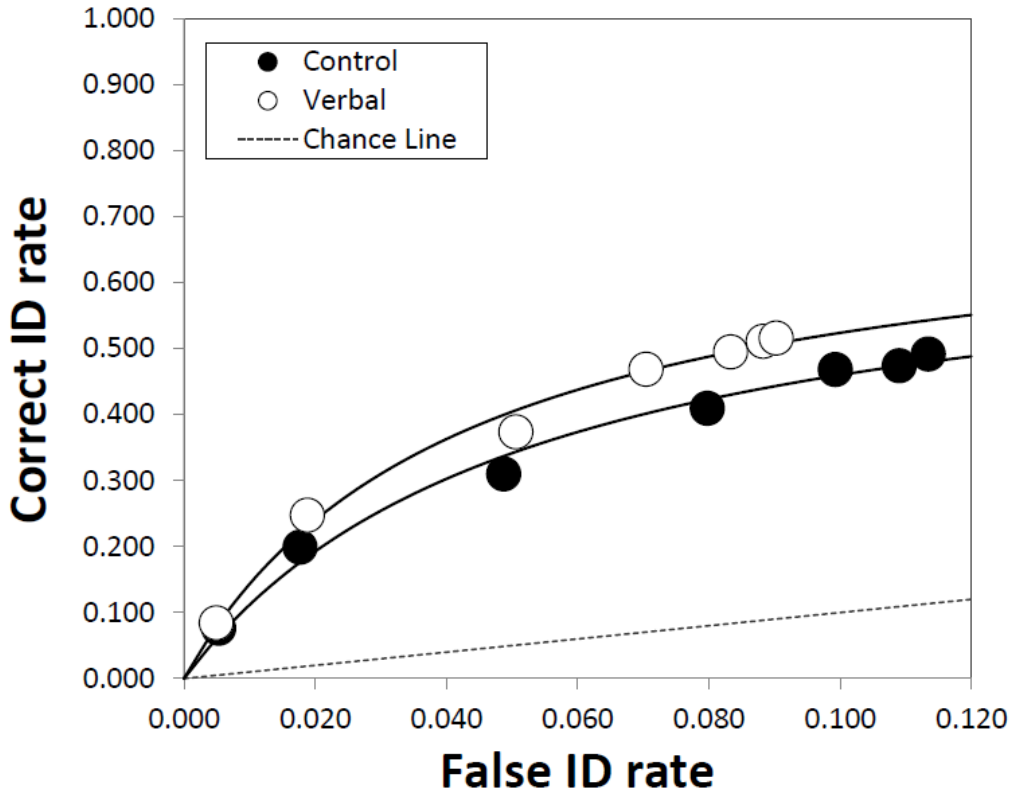

Figure 2. Receiver operating characteristic (ROC) curves for the Verbal and Control conditions. The lines through the ROC curves were estimated using a hyperbolic function. The dashed line represents chance performance.

CAC Analysis

Confidence levels were binned into low (1-3), medium (4-5), and high (6-7) because there were too few responses in some levels. For each level of confidence, suspect ID accuracy (*A*) = # correct suspect IDs / (# correct suspect IDs + # innocent suspect IDs). The CAC curves shown in Figure 3 appear to show the same level of accuracy for low, medium, and high levels of confidence between the two conditions. Note that the Appendix contains the CAC analysis using the other method of estimating the false ID rate. Namely, the most often identified filler was designated as the innocent suspect. The conclusions are the same regardless of which procedure is used.

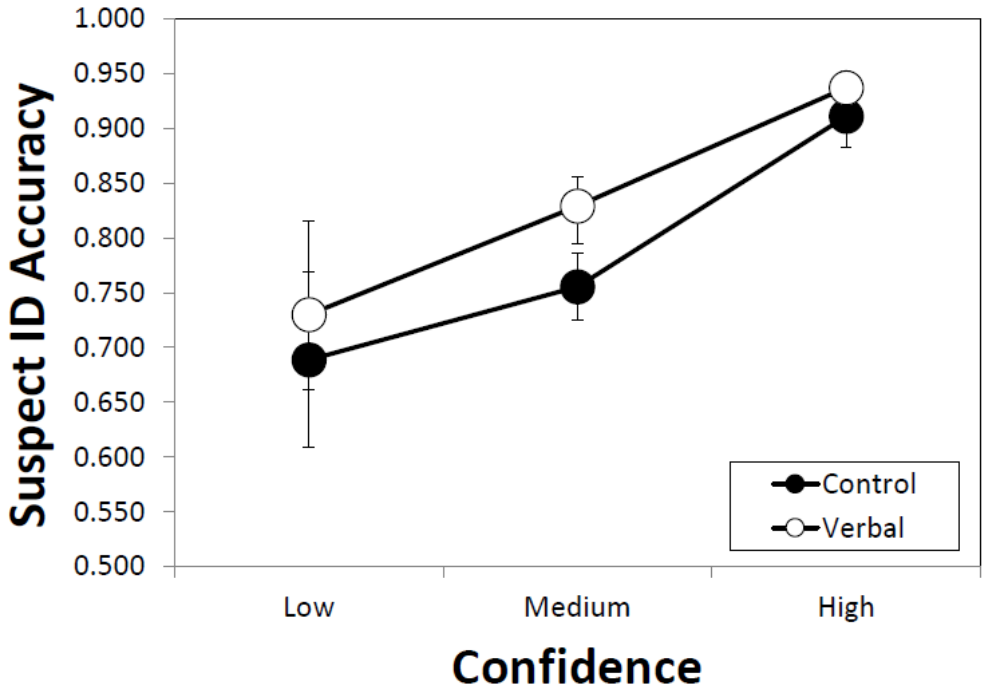

Figure 3. Confidence-accuracy characteristic (CAC) curves for Verbal and Control Condition. The bars represent standard error bars. Note that the error bar for high confidence judgments in the verbal condition is very small.

Discussion

Alogna et al. (2014) attempted a direct replication of Experiment 4 from Schooler and Engstler-Schooler (1990). Participants watched a video of a mock crime and,

immediately after watching the video, either described the perpetrator or completed a control task. Twenty minutes later participants attempted to identify the perpetrator from a perpetrator-present lineup. Alogna et al. found a small, statistically significant reduction in the correct ID rate for participants who had previously described the perpetrator, replicating Schooler and Engstler-Schooler's original result. However, because the false ID rate was not measured in conjunction with the correct ID rate, it is unclear whether this verbal overshadowing effect reflects a reduction in discriminability or just a conservative shift in response bias. Several theories suggest that verbal overshadowing reflects a reduction in discriminability (e.g. Schooler & Engstler-Schooler, 1991; Schooler et al., 1997; Meissner et al., 2001), but more recently Clare and Lewandowsky (2004) have argued that verbal overshadowing reflects a conservative shift in response bias rather than a reduction in discriminability. To determine whether verbalization affected response bias or discriminability, a near-direct replication was conducted, but a perpetrator-absent lineup condition was included in order to measure both correct and false ID rates. With both correct and false ID rates, ROC curves can be constructed and statistically compared in order to reveal any significant difference in discriminability and response bias (Mickes et al., 2012). In this experiment, verbally describing the perpetrator *did not* result in a reduction in discriminability. ROC analysis revealed no significant difference in discriminability between participants who had described the perpetrator and control participants. Below, this finding is discussed with respect to each theory of verbal overshadowing.

Content Accounts of Verbal Overshadowing

The content accounts of verbal overshadowing predict worse discriminability for participants who had described the perpetrator because the description interferes with the memory of the perpetrator (Schooler & Engstler-Schooler, 1990; Meisner et al., 2001). In this experiment, description-related interference should have had a maximum effect on discriminability because participants described the perpetrator immediately after watching the mock crime video. At this time, the memory of the perpetrator was likely undergoing consolidation and was, therefore, highly susceptible to interference. In list-learning

paradigms, for instance, it has long been demonstrated that the effects of interference have the strongest impact on memory performance when the interfering list is presented immediately after the study list; if the interfering list is presented at a later time during the retention interval (i.e. when more of the items on the study list have been consolidated), then the interference has less of a detrimental impact on memory performance (Muller & Pilzeker, 1900; Skaggs, 1925; Postman & Alper, 1946; for review, see Wixted, 2004). Yet, there was no significant difference in discriminability between the verbal and control condition. The fact that participants described the perpetrator immediately after watching the mock crime video and yet, no difference in discriminability was found, strongly suggests that description-related interference is, perhaps, not substantial enough to cause a reduction in discriminability.

Transfer-Inappropriate Processing Account

According to the transfer-inappropriate processing account of verbal overshadowing (Schooler et al., 1997), participants who describe the perpetrator switch to a processing style that makes it difficult for participants to identify the perpetrator from a lineup. Although this account generally predicts reduced discriminability for participants who had described the perpetrator, because there was a twenty minute delay between description and identification, participants may have had enough time to switch back to a processing style more conducive to identification. Chin and Schooler (2008) argued that "time or other peripheral events may easily shift processing back to a more global orientation, thus eliminating the verbal overshadowing effect" (p. 404). This argument, however, is rather unconvincing because it is not clear how the passage of time, per se, causes participants to switch processing styles. It seems just as dubious to say, for example, that the passage of time itself is the cause of forgetting, a point which McGeoch (1932) has articulated long ago. A more reasonable explanation for this finding might actually link the presence or absence of factors during the retention interval to changes in processing style.

Criterion-shift Account of Verbal Overshadowing

Although no significant difference in discriminability was observed, participants who had described the perpetrator were more conservative when making an identification. This finding is consistent with the criterion-shift account of verbal overshadowing (Clare & Lewandowsky, 2004). These participants most likely experienced difficulty during the description task and, as a result, believed their memory for the perpetrator was poor (even though they were just as good at discriminating innocent from guilty suspects as controls). This belief may have caused participants to make an identification only when they felt very confident in the accuracy of their memory (i.e. choosing to reject the lineup when they felt unsure).

Verbal Descriptions and Reliability

To determine whether confidence and accuracy were related for all levels of confidence, CAC analysis was conducted. CAC analysis revealed that identifications made with high confidence were higher in accuracy than identifications made with medium confidence which, in turn, were higher in accuracy than identifications made with low confidence. Furthermore, CAC analysis revealed that identifications made with high confidence were comparably reliable for *both* groups. The ROC result is of theoretical interest, but, perhaps the more informative results are from CAC analysis given that police investigators are unlikely to stop collecting verbal descriptions. Many researchers believe that suspect IDs admitted as evidence in court "…should be weighted less if the witness had provided a description earlier" (Alogna et al., 2014, p. 557). Yet, the CAC curves did not differ whether or not participants provided verbal descriptions. The upshot is that identifications made with high confidence were likely to be accurate even in the verbal condition. That is, those identifications are reliable and if that result replicates, then that is an important finding for judges and jurors.

Experiment 3

After the first registered replication attempt (Alogna et al., 2014), 22 of those laboratories ran another replication (attempting to replicate Experiment 1 from Schooler

& Engstler-Schooler, 1990). This time, participants described the perpetrator or completed a control task for the allotted time immediately before viewing the lineup (see Figure 4). The meta-analytic effect across the 22 studies found a much larger verbal overshadowing effect (i.e. a much larger reduction in the correct ID rate). Still, the same issue applies. It is not clear whether the reduction in correct ID rate reflects a reduction in discriminability or a conservative shift in response bias.

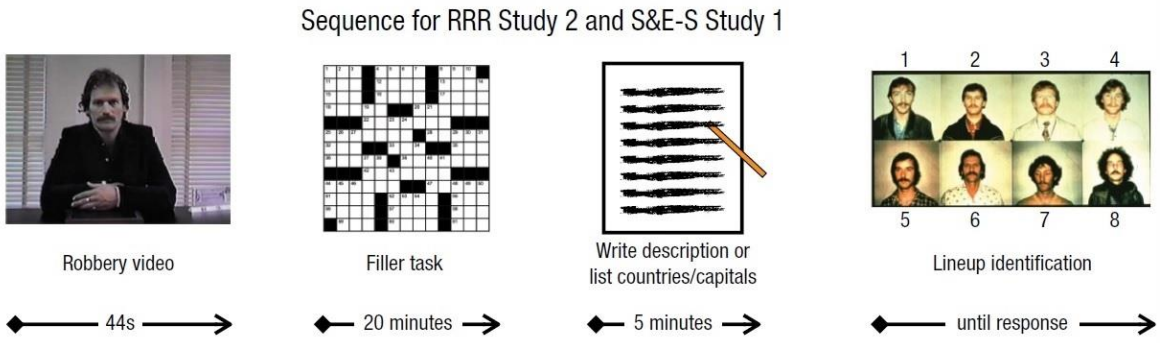

Figure 4. Task sequence for the second registered replication report (RRR Study 2) from Alogna et al. (2014).

Methods

Participants

Participants (N = 780) were recruited from Royal Holloway, University of London (n = 138), Amazon Mechanical Turk (n = 245), and SampleSize (n = 397). The participants (n = 10) who reported previously viewing the video were excluded from the analyses. The remaining participants (n = 770, 442 female; 318 male; 10 did not state; age in years: M = 27.9, SD = 11.1) were randomly assigned to the control condition or to the verbal condition and a perpetrator-absent lineup ($n_{control}$ = 179; n_{verbal} = 185) or a perpetrator-present lineup ($n_{control}$ = 196; n_{verbal} = 210). Royal Holloway, University of London Ethics Board approved this study.

Materials

The materials used throughout the study were the same materials used in Experiment 1 and the original report (Schooler & Engstler-Schooler, 1990).

Procedure

Participants studied a video of a bank robbery, engaged in a 20-minute distractor task, either provided a verbal description of the target (i.e. the bank robber) or completed a control task before viewing a lineup (see Figure 4). The lineup either did or did not contain the perpetrator. The size of the lineup was modified from eight to six members which enabled participants to be randomly assigned to perpetrator-present and perpetrator-absent conditions while keeping the size of the lineup constant. Confidence in the identification decision was also collected in order to conduct ROC and CAC analyses.

Results

Descriptive Analysis

The total number of correct IDs, false IDs, and filler IDs for every level of confidence for the verbal and the control condition are shown in Table 3.

Table 3

Response frequencies are displayed for every decision outcome per level of confidence.

	Verbal Condition					Control Condition				
	Perp Present			Perp Absent		Perp Present			Perp Absent	
Confidence	CID	FID	noID	FID	noID	CID	FID	noID	FID	noID
1	0	1	2	1	0	0	2	0	1	0
2	2	5	3	4	1	3	3	2	4	6
3	5	8	11	13	5	7	5	2	13	7
4	12	10	14	29	17	21	10	9	29	19
5	29	13	34	33	32	47	10	15	33	23
6	30	4	18	10	38	25	3	9	10	20
7	2	2	5	3	15	19	0	4	3	11

Note: Perp, perpetrator; ID, identification; CIDs, correct IDs; FIDs, filler IDs.

ROC Analysis

The suspect ID rates for perpetrator-present lineups (i.e. correct ID rates), suspect ID rates for perpetrator-absent lineups (i.e. false ID rates) and filler ID rates for each level of confidence are shown in Table 4. The italicized values were used to construct the ROC curves shown in Figure 5. The original (Schooler & Engstler-Schooler, 1990) and replicated reports (Alogna et al., 2014) observed a -25% [95% confidence interval: -45% to -5%] and -16% [95% confidence interval: -21% to -1%] reduction in the correct ID rate due to verbalization, respectively. We observed a -24% [95% confidence interval: -33% to -14%] reduction in the correct ID rate. The correct ID rate findings, ignoring false ID rates, replicated both the original and replicated experiments.

Table 4

Correct ID, false ID, filler ID and no ID rates per level of confidence per condition.

		Perp Present			Perp Absent		
	Confidence	Correct ID	Filler ID	noID	False ID	Filler ID	noID
Verbal	1	***.38***	.21		***.07***	.42	
	2	*.38*	.20		*.07*	.41	
	3	*.37*	.18		*.06*	.38	
	4	*.35*	.14	.41	*.06*	.34	.58
	5	*.29*	.09		*.04*	.25	
	6	*.15*	.03		*.02*	.10	
	7	*.01*	.01		*.00*	.02	
Control	1	***.62***	.21		***.09***	.52	
	2	*.62*	.20		*.09*	.51	
	3	*.61*	.18		*.08*	.49	
	4	*.57*	.14	.21	*.07*	.42	.48
	5	*.46*	.09		*.04*	.26	
	6	*.22*	.03		*.01*	.07	
	7	*.10*	.01		*.00*	.02	

Note: Perp, perpetrator; ID, identification.

ROC analysis was conducted to determine whether verbalization affects discriminability and not just response bias. The ROC curves shown in Figure 5 reveal that the verbal condition ROC is lower than the control condition ROC and that difference was significant. The *pAUC* was calculated by using the false ID cutoff of .59 for both conditions. The *pAUC* for the control condition (0.15) was significantly greater than the *pAUC* for the verbal condition (0.09), $D = 3.15$, $p = .002$. Note that the Appendix contains a separate ROC analysis that estimated the false ID rate by taking the most often identified filler as the innocent suspect. The conclusions are the same regardless of which procedure is used.

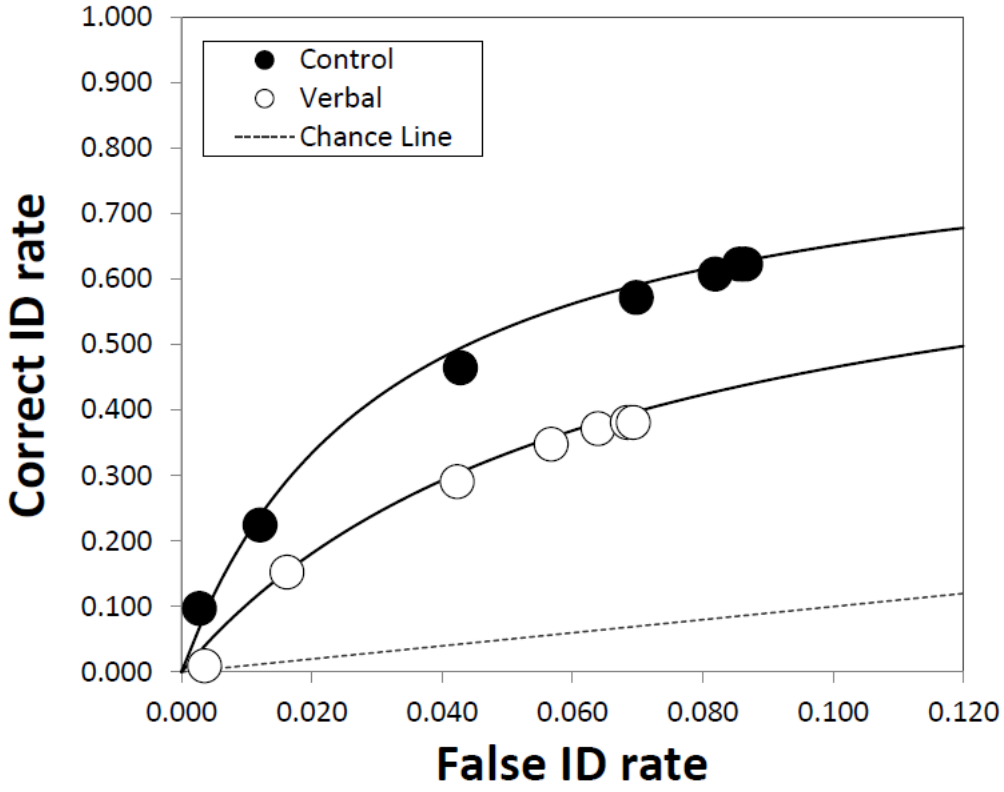

Figure 5. Receiver operating characteristic (ROC) curves for the Verbal and Control conditions. Best-fitting lines are drawn through the curves. The dashed line represents chance performance.

CAC Analysis

Confidence levels have been binned into low (1-3), medium (4-5), and high (6-7) because there were too few responses in some levels. For each level of confidence, suspect ID accuracy (*A*) = # correct suspect IDs / (# correct suspect IDs + # innocent suspect IDs). The CAC curves in Figure 6 show similar accuracy for each level of confidence regardless of condition. Although discriminability is lower in the verbal condition (as determined by ROC analysis), suspect IDs made in the verbal condition are as reliable as the suspect IDs made in the control condition. The Appendix contains a separate CAC analysis that estimated the false ID rate by taking the most often identified filler as the innocent suspect. The conclusions are the same regardless of which method is used.

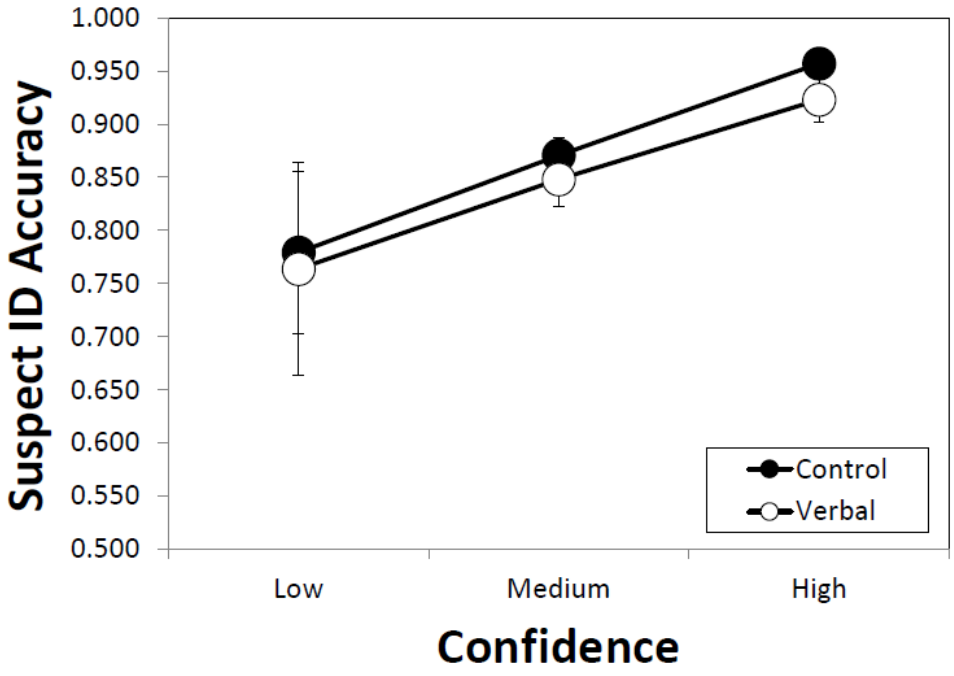

Figure *6.* Confidence-accuracy characteristic (CAC) curves for Verbal and Control conditions. The bars represent standard error bars.

Discussion

Alogna et al. (2014) attempted to directly replicate Experiment 1 from Schooler and Engstler-Schooler (1990). Participants watched a video of a mock crime and engaged in a twenty minute distractor task. Participants then either described the perpetrator or

completed a control task right before attempting to identify the perpetrator from a perpetrator-present lineup. Thus, Experiment 2 was the same as Experiment 1 with the exception that the description was provided right before identification (whereas the description was provided twenty minutes prior to identification in Experiment 1). Alogna et al. found a large, statistically significant reduction in the correct ID rate for participants who had previously described the perpetrator, replicating Schooler and Engstler-Schooler's original result. However, this result may either reflect a reduction in discriminability or a conservative shift in response bias. A near-direct replication was conducted that included a perpetrator-absent lineup condition in order to measure correct *and* false ID rates and conduct ROC analysis (Mickes et al., 2012).

The results from ROC analysis indicate that participants who had described the perpetrator immediately before identification had significantly worse discriminability than control participants. Yet, in Experiment 3, when participants in the verbal condition provided descriptions twenty minutes prior to identification, discriminability did not differ from the control condition. Why is there no significant difference in discriminability when the delay between description and identification is large, but there is a difference in discriminability when the delay is short?

The diagnostic feature-detection (DFD) hypothesis may provide insight into this difference (Wixted & Mickes, 2014). The hypothesis was initially proposed to account for the discriminability advantage that simultaneous lineup presentations have over procedures that involve showing an individual in isolation (as with sequential lineups or showups). By seeing the lineup members together, it is readily apparent to the eyewitness that there are features shared across lineup members that should be discounted because they are not diagnostic of guilt. For example, if the perpetrator were a young, White male, then adding weight to those features would not be helpful because all of the lineup members would be young, White males. As a result, when eyewitnesses see the lineup member together, they are able to discount features that are shared, and thus focus on features that are diagnostic of guilt. This strategy is less likely to be used when lineup members are presented individually. The same concept may help to explain why verbal

descriptions only impair discriminability when they are made after a delay. More specifically, participants may use more diagnostic feature descriptions immediately after encoding than they do after a delay. After a delay, by contrast, the description may become more general, corresponding to the common features that everyone in the lineup shares. In that case, the participants may have a tendency to rely on the description they just gave when trying to identify the face of the perpetrator. To the extent that they rely on the general (shared) facial features mentioned in the verbal description, discriminability would be impaired. To assess whether or not the DFD hypothesis can help to account for the differences in discriminability when verbal descriptions are delayed, a content analysis of the verbal descriptions provided in Experiments 1 and in Experiment 2 was conducted.

Content Analysis

To conduct the content analysis, 20 words were identified based on the appearance of the eight images of the perpetrator and fillers. Ten words were selected that would likely be useful in differentiating the perpetrator from fillers (diagnostic-feature words) and 10 words were selected that would not likely be useful in differentiating the perpetrator from fillers (non-diagnostic-feature words). The diagnostic-feature words were descriptors that were not shared by all of the lineup members. The diagnostic-feature words that were selected were *chin*, *jaw*, *cheek*, *brow*, *forehead*, *eye*, *oval*, *round*, *wavy*, and *point*. The non-diagnostic-feature words were descriptors that are shared by all of the lineup members. The non-diagnostic-feature words we selected were *white*, *male*, *age*, *brown*, *black*, *moustache*, *dark*, *weight*, *build*, and *height*. The diagnostic-feature and non-diagnostic feature words were counted from descriptions provided by participants in the verbal condition in Experiment 1 (immediate descriptions) and compared with those descriptions in Experiment 2 (delayed descriptions).

Significantly more diagnostic feature words were used when verbal descriptions were provided immediately after watching the mock crime video (Experiment 1; $M = 2.3$, $SD = 1.5$) compared to when verbal descriptions were provided 20 minutes after watching the mock crime video (Experiment 2; $M = 2.0$, $SD = 1.3$, $t(785) = 2.57$, $p = .01$). Those who described the perpetrator immediately after watching the mock crime video

(Experiment 1) also provided fewer non-diagnostic feature words ($M = 2.9$, $SD = 1.6$) than those who described the perpetrator 20 minutes later (Experiment 2; $M = 3.6$, $SD = 1.7$, $t(785) = 5.03$, $p < .001$). A 2 × 2 analysis of variance revealed a significant interaction between type of feature (diagnostic vs. non-diagnostic) and time of verbal description (immediate vs. delayed), $F(1, 1570) = 30.2$, $p < .001$. These results provide evidence for the DFD hypothesis.

Experiment 5

According to the DFD hypothesis (Wixted & Mickes, 2014), the verbal overshadowing effect may be due to the fact that participants tend to describe non-diagnostic features, which are features shared by everyone in the lineup. To the extent that non-diagnostic features are given weight, participants will be less able to discriminate the perpetrator from a similar-looking member in the lineup. An analysis of the written descriptions provided by participants in Experiments 3 and 4 showed a difference in the type of features used to describe the perpetrator. Participants in Experiment 3 described more diagnostic features of the perpetrator. For instance, participants described the perpetrator's *cheek*, *jaw*, *brow*, *eyes*, and *forehead* more so than participants in Experiment 4. These participants had greater discriminability than control participants, though the difference was not significant. Participants in Experiment 4 described more non-diagnostic features of the perpetrator such as the perpetrator's *ethnicity*, *gender*, *age*, *weight*, and *height*. Consistent with the DFD hypothesis, these participants had significantly worse discriminability than control participants.

These results indicate that the amount of diagnostic and non-diagnostic words used to describe the perpetrator may have an impact on discriminability. It is difficult to make any firm conclusions, however, because these two groups of participants were compared across experiments. In Experiment 5, participants watched a video of a mock crime and were instructed to describe either the specific features of the perpetrator, the general features of the perpetrator, or a set of items that were unrelated to the crime video before viewing a perpetrator-present or perpetrator-absent lineup. By instructing participants to describe the specific or general features of the perpetrator, we sought to manipulate the

amount of diagnostic and non-diagnostic words used to describe the perpetrator. We reasoned that the features specific to the perpetrator were less likely to be shared amongst the fillers in the lineup and so, these features were much more likely to be diagnostic of guilt. Whereas, the general features of the perpetrator were much more likely to be shared amongst the fillers in the lineup. Thus, these features were much more likely to be non-diagnostic of guilt. Confidence in the identification decision was collected in order to conduct ROC and CAC analyses. It was predicted that those in the general description condition would use fewer diagnostic and more non-diagnostic words to describe the perpetrator than those in the specific description condition and, as a result, would yield worse discriminability. However, although there may be a difference in discriminability between groups, both groups of participants may be equally reliable.

Methods

Participants

Participants (N = 948) were recruited from Amazon Mechanical Turk. Participants that failed to answer the validation question correctly (stating the crime committed; n = *23*) were excluded from all analyses. The remaining (n = 925; 399 male, 523 female, 3 unspecified; age in years: M = 31.8, SD = 10.7) participants were randomly assigned to either the unrelated group (n = 310), the specific feature group (n = 310), or the general feature group (n = 305). Participants were then randomly assigned to be tested on either a perpetrator-present lineup ($n_{unrelated}$ = 158; $n_{specific}$ = 156; $n_{general}$ = 144) or a perpetrator-absent lineup ($n_{unrelated}$ = 152; $n_{specific}$ = 154; $n_{general}$ = 161). Royal Holloway, University of London Ethics Board approved this study.

Materials

The crime video featured a young white male walking into a lobby where an unattended laptop was located. The perpetrator walks down the hall towards the camera, quickly surveys the area and steals the laptop. The perpetrator's face was shown for roughly 12 seconds. A front-face photograph of the perpetrator's face was used in all of the perpetrator-present lineups. All of the fillers were young, white males that matched

the description of the perpetrator. Fillers were selected from the Florida Department of Corrections Offender Network (http://www.dc.state.fl.us/AppCommon/) using the following search criteria: male, 20-21 years old, white, height of 5'10"-6'2", brown or black, short hair with no facial hair and no distinguishing features. From here, 100 faces that matched these descriptions were collected and grey-scaled. For each participant, five filler images were randomly retrieved for a perpetrator-present lineup and six filler images were randomly retrieved for a perpetrator-absent lineup. Thus, there was no designated innocent suspect placed in perpetrator-absent lineups.

Procedure

After consenting to participate, participants studied a video of a mock crime and were then instructed to either describe the specific features of the perpetrator, the general features of the perpetrator, or describe a set of items that were unrelated to the crime video. After engaging in a 5-minute distractor task, participants were then presented a six-person lineup that either did or did not contain the perpetrator. Confidence in the identification decision was collected using an 11-point scale ranging from 0% (just guessing) to 100% (absolutely certain). Confidence was collected in order to conduct ROC and CAC analyses. After the lineup, participants answered a set of questions about the video including a validation question ("What crime was committed?") before being debriefed.

Results

Descriptive Analysis

There were 23 participants who failed to answer the validation question correctly and were therefore excluded from all analyses. The total number of perpetrator-present and perpetrator-absent trials as well as the number of correct IDs, false IDs, and filler IDs for every level of confidence for the general feature, specific feature, and unrelated conditions are shown in Table 5. A one-way analysis of variance was conducted to determine whether the length of the descriptions was significantly different among the three groups of participants. This analysis revealed a significant difference in the length of descriptions among the three groups of participants, $F(2,924) = 242.6$, $p < .001$).

Subsequent t-tests revealed that participants assigned to the unrelated condition provided longer descriptions (M = 67.7, SD = 38.4) than those assigned to the specific feature condition (M = 34.2, SD = 21.0, $t(613) = 13.4$, $p < .001$), and those assigned to the specific feature condition provided longer descriptions than those assigned to the general feature condition (M = 22.0, SD = 14.8, $t(618) = 8.39$, $p < .001$). These differences might have affected the ROC results discussed below.

Table 5

Response frequencies for every identification decision outcome are displayed for each level of confidence for Unrelated, Specific Feature, and General Feature conditions.

		Perp Present			Perp Absent	
		Correct IDs	Filler IDs	no IDs	Filler IDs	no IDs
Unrelated	0	0	2	5	7	0
	10	0	0	1	3	1
	20	1	3	1	0	1
	30	1	7	5	9	2
	40	0	4	4	16	7
	50	4	10	7	19	11
	60	3	3	9	12	10
	70	4	11	5	9	15
	80	8	6	16	9	11
	90	5	3	8	3	9
	100	4	0	4	3	4
Specific	0	0	0	2	2	4
	10	0	1	1	1	0
	20	0	3	4	5	1
	30	3	3	7	5	3
	40	1	4	3	6	6
	50	4	8	8	12	8
	60	7	9	8	6	10
	70	7	9	19	11	16

		Perp Present			Perp Absent	
		Correct IDs	Filler IDs	no IDs	Filler IDs	no IDs
	80	8	3	7	9	16
	90	6	1	11	3	16
	100	1	1	7	1	13
General	0	0	3	2	1	2
	10	0	0	1	0	1
	20	1	3	2	2	1
	30	1	4	6	8	1
	40	4	4	2	7	1
	50	6	6	10	10	12
	60	8	4	10	12	9
	70	9	7	9	7	23
	80	9	4	10	4	18
	90	6	2	10	2	15
	100	3	1	11	3	13

Note: Perp, perpetrator; ID, identification.

ROC Analysis

The suspect ID rates for perpetrator-present lineups (i.e. correct ID rates), suspect ID rates for perpetrator-absent lineups (i.e. false ID rates) and filler ID rates for both perpetrator-present and perpetrator-absent lineups for each level of confidence are shown in Table 6. The bold italicized values are the overall correct and false ID rates. The correct ID rate in the general feature condition (.297) was greater than the correct ID rates in the specific feature condition (.237) and the unrelated condition (.208). The false ID rate in the unrelated condition (.093) was greater than the false ID rates in the specific condition (.066) and the general condition (.061). The unrelated condition yielded the lowest correct ID rate (.208) and the highest false ID rate (.093), which suggests that discriminability was worse in this condition than in the general or specific feature conditions. The specific feature condition yielded a lower correct ID rate and higher false ID rate than the general

feature condition, which also suggests that discriminability was worse for this condition compared to the general feature condition. In order to determine whether discriminability was significantly different across the three conditions, the correct and false ID rates for each level of confidence were plotted in ROC space in order to construct ROCs for each condition. The italicized values in Table 6 were used to construct the ROC curves shown in Figure 7.

Table 6

Correct ID, false ID, filler ID and no ID rates for perpetrator-present and perpetrator-absent lineups by level of confidence per condition.

		Perp Present			Perp Absent		
	Confidence	CID	Filler ID	noID	FID	Filler ID	noID
Unrelated	0	***.21***	.34		***.10***	.62	
	10	*.21*	.34		*.10*	.58	
	20	*.21*	.33		*.09*	.56	
	30	*.20*	.31		*.08*	.50	
	40	*.20*	.26		*.07*	.44	
	50	*.19*	.23	.45	*.06*	.34	.44
	60	*.17*	.16		*.04*	.22	
	70	*.15*	.14		*.03*	.15	
	80	*.12*	.06		*.02*	.09	
	90	*.06*	.02		*.01*	.04	
	100	*.03*	.02		*.00*	.02	
Specific	0	***.24***	.27		***.07***	.40	
	10	*.24*	.27		*.06*	.38	
	20	*.24*	.26		*.06*	.38	
	30	*.24*	.24		*.06*	.34	
	40	*.22*	.22		*.05*	.31	
	50	*.21*	.20	.49	*.05*	.27	.60
	60	*.19*	.15		*.03*	.20	
	70	*.14*	.09		*.03*	.16	

		Perp Present			Perp Absent		
	Confidence	CID	Filler ID	noID	FID	Filler ID	noID
	80	*.10*	.03		*.01*	.08	
	90	*.05*	.01		*.00*	.03	
	100	*.01*	.01		*.00*	.01	
General	0	***.30***	.24		***.06***	.37	
	10	*.30*	.22		*.06*		
	20	*.30*	.22		*.06*	.36	
	30	*.29*	.20		*.06*	.35	
	40	*.29*	.18		*.05*	.30	
	50	*.26*	.15	.46	*.04*	.25	.63
	60	*.22*	.11		*.03*	.18	
	70	*.17*	.10		*.02*	.11	
	80	*.11*	.04		*.01*	.06	
	90	*.06*	.02		*.01*	.03	
	100	*.02*	.01		*.00*	.02	

Note: Perp, perpetrator; ID, identification; FID, false IDs.

The ROC curves shown in Figure 7 reveal that the general feature condition ROC is greater than the specific feature condition ROC which is greater than the unrelated condition ROC. The *pAUC* for all three conditions was calculated by using the false ID cutoff of .939. Because three comparisons were being made, the alpha level was corrected using Bonferroni corrections so that a significant difference would be accepted as $p < .017$. The *pAUC* from the general feature condition and the unrelated condition ROCs were first compared. The *pAUC* for the general feature condition (0.012) was not significantly greater than the *pAUC* for the unrelated condition (0.007), $D = 1.28$, $p = .23$. Next, the *pAUC*s from the general feature condition and specific feature ROC were compared. Again, the *pAUC* was calculated by using the false ID cutoff of .939. The *pAUC* for the general feature condition (0.012) was not significantly greater than the *pAUC* for the unrelated condition (0.008), $D = .783$, $p = .43$. Lastly, the *pAUC*s from the specific feature condition and the unrelated condition was compared. Although the *pAUC* for the specific

feature condition (.008) was greater than the *pAUC* for the unrelated condition (.007), there was no significant difference, $D = .424$, $p = .67$.

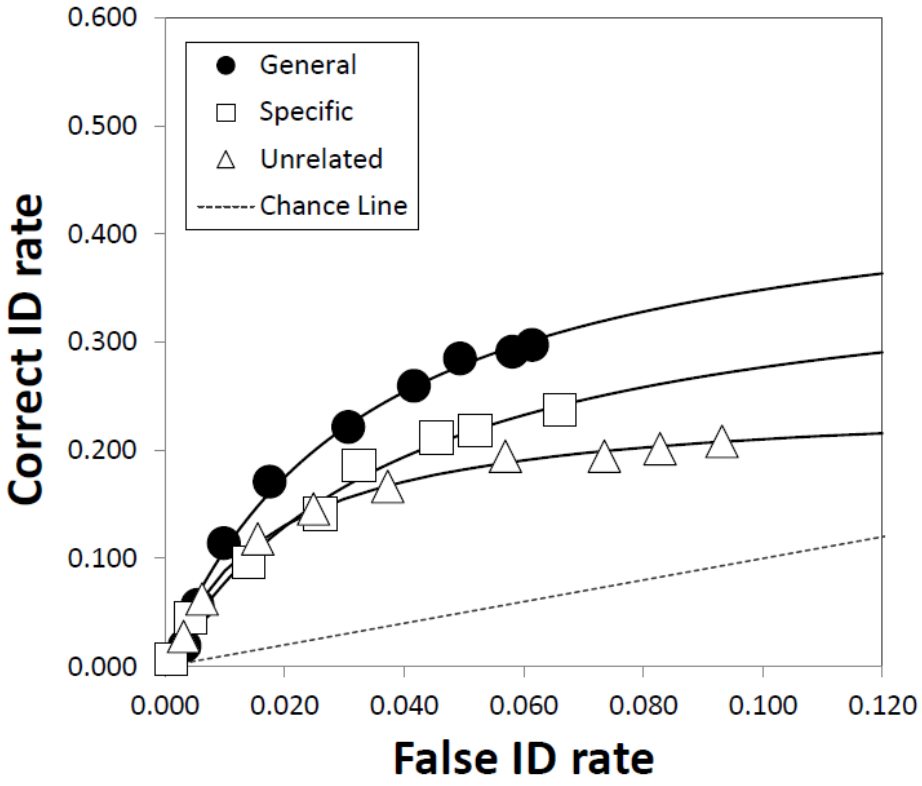

Figure 7. Receiver operating characteristic (ROC) curves for the General, Specific, and Unrelated conditions. The lines through the ROC curves were estimated using a hyperbolic function. The dashed line represents chance performance.

CAC Analysis

Confidence levels have been binned into low (0% – 60%), medium (70% – 80%), and high (90% – 100%) because there were too few responses in some levels. For each level of confidence, suspect ID accuracy (*A*) = # correct suspect IDs / (# correct suspect IDs + # innocent suspect IDs). The CAC curves in Figure 8 show similar accuracy for medium and high levels of confidence regardless of condition. This means that suspect IDs, with the exception of low confident suspect IDs, made in the general or specific feature condition are as reliable as the suspect IDs made in the unrelated condition.

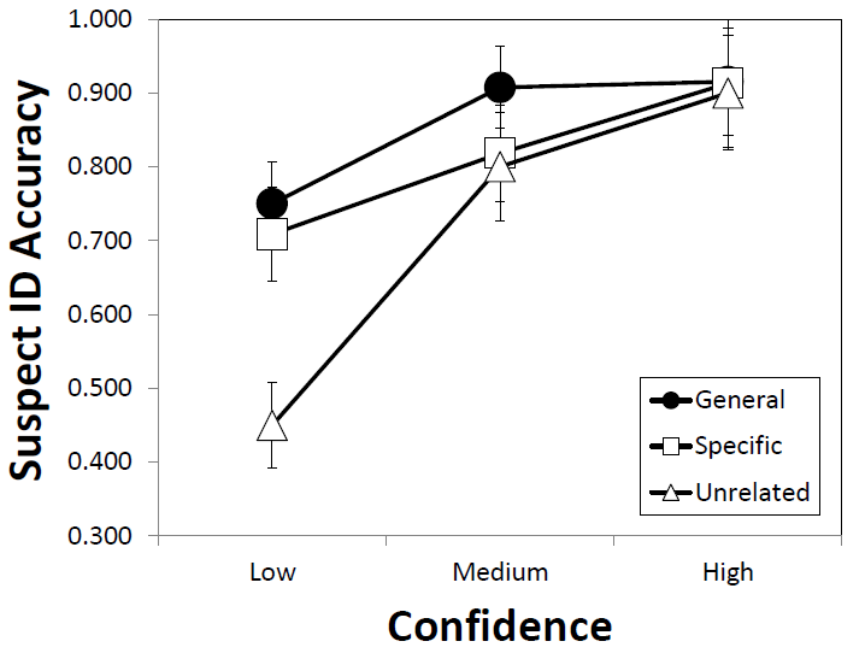

Figure 8. Confidence-accuracy characteristic (CAC) curves for the General, Specific, and Unrelated conditions. The bars represent standard error bars.

Discussion

In this experiment, we manipulated description instructions to encourage the use of diagnostic or non-diagnostic descriptions. We reasoned that the amount of diagnostic and non-diagnostic words used to describe the perpetrator could affect discriminability or reliability from a lineup. Participants watched a video of a mock crime and either described the specific features of the perpetrator, the general features of the perpetrator, or described several unrelated items. ROC analysis revealed no significant difference in discriminability among the three groups of participants. Although participants described the perpetrator using specific or general features, participants in the unrelated condition wrote, on average, more words than participants in the general and specific feature conditions. In fact, participants in the general and specific feature conditions hardly used any words to describe the perpetrator beyond the basic requirements, perhaps because the task was too difficult. If participants wrote more details about the general or specific features of the perpetrator such that the length of the descriptions were the same as the

length of descriptions in the unrelated condition, then a difference in discriminability might have been observed.

In order to determine whether describing the perpetrator impacted the reliability of participants' suspect IDs, CAC analysis was conducted. CAC analysis revealed that confidence and accuracy are, once again, related. Specifically, participants who expressed low confidence in their suspect IDs were less accurate than those who expressed medium confidence who, in turn, were less accurate than those who expressed high confidence. Across medium and high confidence, there appeared to be no significant difference in suspect ID accuracy across the three conditions. Thus, despite participants describing the perpetrator, which has been shown to impact discriminability (e.g. in Experiment 2), they were just as reliable across medium and high confidence as those who had not previously described the perpetrator. This result is informative to judges and jurors who may need to determine the reliability of a suspect ID that was made after an eyewitness described the perpetrator to the police.

Chapter 5

The Effect of a Weapon on Discriminability and Reliability

If a perpetrator is armed during a crime, an eyewitness may focus on the weapon the perpetrator is holding at the expense of focusing on the perpetrator's face (Loftus, Loftus, & Messo, 1987). The "weapon focus" effect is expected to impact memory in two ways: 1) the eyewitness is less able to accurately recall contextual details of the crime and 2) the eyewitness is less able to discriminate guilty from innocent suspects in a lineup. The focus of this chapter is to determine to what extent weapon focus impacts discriminability from a lineup and to address the reasons why this effect might occur. Throughout the last thirty years, two theories have been tested.

Weapon Focus in the Laboratory

Arousal/Stress Hypothesis

In a real-life crime situation, a perpetrator holding a weapon likely induces considerable emotional stress in an eyewitness. Past research has shown that the level of emotional stress or arousal a participant experiences is closely related to how that participant performs on a variety of tasks; too much or too little arousal can hinder performance (e.g. Yerkes & Dodson, 1908; Hebb, 1955; Broadhurst, 1959; Duffy, 1962; Humphreys & Revelle, 1984). Easterbrook (1959) proposed that attention and arousal interact to determine which aspects of the environment are utilized. According to this theory, high levels of arousal could lead to a perceptual or attentional narrowing causing the source of the arousal to be over-monitored and other peripheral aspects of the environment to be unmonitored and underutilized.

Based on Easterbrook's (1959) hypothesis, Loftus (1979) argued that the threat of a weapon elicits high levels of physiological arousal which could cause eyewitnesses to focus almost exclusively on the weapon (the source of the arousal), at the expense of attending to the peripheral details, such as the perpetrator's face. Loftus et al. (1987) attempted to demonstrate this in the laboratory. In Experiment 1, participants studied a set

of slides which either depicted a customer (i.e. the target) paying for a meal at the counter of a restaurant with a check or depicted the target revealing a gun and demanding money from the cashier. The two series of slides were virtually identical except that in one condition the target hands a check to the cashier (weapon-absent condition) and, in the other condition, the target points a gun at the cashier (weapon-present condition). The target's face was in clear view in both conditions. Following a 15 minute distractor task, participants attempted to identify the target from a 12-person lineup. The target was always present in the lineup. Those in the weapon-present condition were less likely to correctly identify the target than those in the weapon-absent condition, although the effect was only marginally significant. This marginal effect was likely due to a small sample of participants. In Experiment 2, the same procedure from Experiment 1 was used, but twice as many participants were recruited. In this experiment, a significant reduction in the correct ID rate was found for those in the weapon-present condition compared to those in the weapon-absent condition. Eye tracking data were also analyzed. Participants fixated more, and for a longer duration, on the gun than they did on the check. This finding is consistent with the arousal/stress hypothesis as the participants fixated for a longer duration on the threatening stimulus (i.e. the gun) than they did on the non-threatening stimulus (i.e. the check) and were less able to correctly identify the perpetrator from a perpetrator-present lineup as a result.

Do Participants Feel Threatened in the Laboratory?

Psychologists have questioned whether a perpetrator shown holding a weapon in a series of images, a video recording, or a staged crime can cause participants to feel personally threatened (e.g. Deffenbacher, 1983). Loftus et al. (1987) did not measure emotional arousal and so, it is difficult to pinpoint whether participants focused on the weapon in response to feeling threatened or for some other reason. Of course, participants in laboratory experiments cannot be threatened with actual guns or knives due to ethical concerns. For this reason, Maass and Kohnken (1989) attempted to simulate the arousal an eyewitness experiences by threatening participants with an injection. Maass and Kohnken told participants that the purpose of the experiment was to investigate the

relation between sport-related physical activity and psychological well-being. This cover story justified the presence of a syringe used for injections in the experimental room. However, the true purpose of the experiment was to determine whether exposure to the syringe, and the threat of an injection in particular, could affect participants' memory for the nurse administering the injection. Simply being exposed to the syringe might cause participants to feel emotionally aroused and cause memory impairment, but Easterbrook's (1959) hypothesis predicts greatest memory impairment for participants who are exposed to the syringe and who feel personally threatened (i.e. those who are threatened with an injection).

After participants were briefed on the cover story, a confederate nurse (i.e. the target) walked in either holding a syringe partially filled with some yellow liquid or a pen. Both the syringe and the pen were held about one meter away from the participant. The nurse then either informed the participant that they would receive an injection (i.e. threatening condition) or informed the participant that they were simply picking up a drug to be given to participants in another room (i.e. non-threatening condition). The nurse then put the syringe or the pen down on the table in front of the participant and left the room. After one minute, an experimenter came in the room and told the participant that they were not going to receive an injection and asked the participant to fill out a questionnaire regarding their mood. Participants then engaged in various distractor tasks for roughly 20 minutes before attempting to identify the nurse from a seven person lineup, but the nurse was *never* present in the lineup.

Exposure to the syringe caused participants' mood to become more negative compared to those exposed to the pen. These participants reported feeling angrier, more agitated, and more nervous. Participants who were exposed to the syringe were also significantly more likely to make a false ID than participants who had been exposed to the pen. Yet, the threat of an injection did not significantly worsen participants' mood and did not lead to a significant increase in false IDs. That is, threatened participants performed just as well as those who were not threatened. These results suggest that the weapon focus effect is at least, in part, an attentional phenomenon. The presence of the syringe (i.e. the

weapon) seemed to grab attention, resulting in worse memory for the target. However, the expected interaction between attention and emotional arousal was not found. It could be that the presence of the weapon was so threatening that the additional threat manipulation was largely ineffective (i.e. a ceiling effect). It could also be that threatening participants with an injection was not threatening enough to increase participants' emotional arousal (i.e. a floor effect). Though, perhaps the most likely explanation is that the interaction between emotional arousal and attentional focus is not driving the weapon focus effect. Similar weapon focus studies in the laboratory have attempted to elicit greater emotional arousal by threatening participants (e.g. Cutler et al., 1987a; Kramer, Buckhout, & Eugino, 1990; Hulse & Memon, 2006). These studies have also failed to find a significant difference in identification accuracy between participants who were threatened and those who were not threatened. This means that something other than emotional arousal may be causing the effect.

Unusual Item Hypothesis

On April 24, 1997, a man walked into a doughnut shop in downtown Toronto and threatened to kill a hostage unless he was given some money. The robber had his arm wrapped around the hostage's neck so tightly that the hostage would soon choke to death. Eyewitnesses to the robbery eventually gave in to the robber's demands and, after receiving a small sum of money, the hostage was released (Mitchell, Livosky, & Mather, 1998). This robbery is notable because the hostage was not a customer or a staff member, but rather a goose from a nearby pond. The situation was so baffling and unexpected to eyewitnesses that they spent most of their time inspecting the goose rather than focusing on the perpetrator's face.

The story of the "goose robber" illustrates how the presence of an unusual or unexpected object can direct attention towards the object and away from the peripheral details of a scene which, in this case, resulted in poorer memory for the perpetrator. Loftus and Mackworth (1978) demonstrated this by showing participants a series of images that contained congruent or incongruent objects. For example, one scene depicted a farmhouse complete with a barn and picket fence, but in the centre of the scene was either a tractor

(congruent object) or an octopus (incongruent object). They found that participants fixated earlier, more often, and for a longer duration on the incongruent objects than the congruent objects. Loftus and Mackworth developed a three stage model of early visual scene processing to explain these findings. According to this model, the general qualities of the scene are analysed first in order to generate the appropriate schema. Once a schema has been activated, the degree to which each object is congruent with the schema is assessed. Any incongruity between the activated schema and objects contained within the scene is resolved by either updating the schema or reinterpreting the object. In order to update the schema or reinterpret the object, attention is directed towards that object and away from the peripheral objects of the scene. Thus, attention is directed towards objects that are incongruent with the schema and away from the peripheral objects of the scene, which could impact memory for those peripheral objects.

Loftus et al. (1987) applied this concept to the weapon focus effect. As previously discussed, they presented a series of images depicting a person standing in front of a cashier inside a restaurant. The person either presented a check to the cashier, as is typically done to pay for a meal, or presented a gun. People do not typically see weapons such as guns inside restaurants and so, the gun was an object that was incongruent with the restaurant schema. Because the gun was unexpected and the check was not, participants spent more time looking at the gun than the check and were also less able to correctly identify the perpetrator.

Can Other Unusual Items Cause Weapon Focus?

If participants focused on the weapon largely because it was unusual or unexpected, then perhaps other unexpected objects can cause a weapon focus effect. Pickel (1998) had participants watch a video of a mock crime where the perpetrator either presented an unusual or a common object. In Experiment 1, a man inside an office walked up to a counter and extended his hand towards a receptionist. In the unusual object condition, the man either held a raw, whole chicken or a gun. These objects were chosen because they were objects not typically seen in the workplace. In the common object condition, the man either held a wallet or a pair of scissors. These objects were chosen because they were

objects one might typically see in the workplace. After the video, participants filled out various questionnaires regarding the perpetrator and other details in the video. Participants then attempted to identify the perpetrator from a perpetrator-present lineup. Participants who were exposed to the unusual objects (i.e. the raw chicken and gun) were less able to recall details about the perpetrator than participants who were exposed to the common objects (i.e. the wallet and scissors), but there was no significant difference in the correct ID rate between participants in the unusual object condition and participants in the common object condition. The same procedure was used in Experiment 2, but the perpetrator in these videos held a different set of objects. The video showed a man walking towards a receptionist in an electronic store. In the unusual object condition, the perpetrator held a Pillsbury doughboy figurine or a butcher knife. In the common object condition, the perpetrator held a screwdriver or a pair of sunglasses. Again, participants that watched the videos of the man holding the unusual objects were less able to recall crime-related details, but there was no significant difference in the correct ID rate between the two groups. In a follow-up study, Pickel (1999) again found no significant difference in the correct ID rate between participants assigned to the unusual object condition and the common object condition.

More recently, Erickson, Lampinen, and Leding (2014) presented participants a series of images depicting a man in a bar carrying either a gun, a rubber chicken, or an empty glass. Participants then attempted to identify the perpetrator from a perpetrator-present or perpetrator-absent lineup. Those assigned to the weapon present condition (i.e. the gun) or the unusual item condition (i.e. the rubber chicken) correctly identified the perpetrator as often as those who were assigned to the normal item condition (i.e. who saw the empty glass). This replicates the previous findings from Pickel (1998; 1999). However, these participants were more likely to falsely identify an innocent suspect compared to participants who were assigned to the normal item condition. In order to determine whether this difference in the false ID rate impacted discriminability, d' was computed for each condition and statistically compared using the G statistic (see Chapter 2 for review). Those in the weapon present condition ($d' = .37$) were significantly worse at discriminating innocent from guilty suspects than those in the normal item condition (d'

= .89), $G = 2.92$, $p < .01$. Yet, what is of interest is that the rubber chicken also reduced discriminability for the perpetrator ($d' = .44$), $G = 2.55$, $p = .01$, consistent with the unusual item hypothesis. However, it is important to note that when this analysis was limited to participants who saw the rubber chicken and the perpetrator in the same image (i.e. the "during" condition), there was no significant difference in discriminability for those in the unusual item condition ($d' = .23$) compared to those in the normal item condition ($d' = .67$), $G = 1.44$, $p = .07$.

Weapon Focus for Actual Crimes

According to Cutshall and Yuille (1989) the events typically seen in the laboratory are not comparable to the events typically seen in actual crimes. They argue that participants in laboratories are essentially uninvolved bystanders who are rarely threatened and rarely feel a personal threat to the extent that an actual eyewitness to a crime is likely to feel. Some have argued, in fact, that the majority of the extant weapon focus findings suffer from weak ecological validity (e.g. Cooper, Kennedy, Herve, and Yuille, 2002). This has encouraged researchers to find evidence of a weapon focus effect outside the confines of a laboratory and within the field (e.g. Cooper et al., 2002). Because it is not known whether the suspect of a crime is factually innocent or guilty, it is difficult to observe whether the presence of a weapon during a crime affects either the correct or false ID rate. However, a weapon focus effect *might* result in a reduction in the suspect ID rate. If memory is worse for the perpetrator because the perpetrator was carrying a weapon, then an eyewitness who was previously exposed or threatened with a weapon might be less likely to identify the suspect and more likely to identify a filler. This pattern could provide some support for the weapon focus effect.

Pike, Brace, and Kynan (2002) analyzed 2,628 live lineups from nine police forces throughout England and Wales. They found that neither the presence of a weapon nor the threat of violence impacted the rate of suspect IDs. Valentine, Pickering, and Darling (2003) analyzed 640 identification decisions from 314 video lineups constructed by the Metropolitan Police in London. Eyewitnesses exposed or threatened with a weapon identified the suspect as often as those eyewitnesses who were not exposed or threatened

with a weapon. Together, the findings in the laboratory and the findings in the field suggest that the presence of a weapon has, at most, a marginal impact on identification accuracy.

Expert Opinion

Although the effect has not often been found in the laboratory (e.g., Cutler & Penrod, 1988; Cutler et al., 1986; Kramer et al., 1990; Pickel, 1998; 1999; Hulse & Memon, 2006; Carlson et al., 2016) and has not been found for actual crimes (Tollestrup, Turtle, & Yuille, 1994; Behrman & Davey, 2001; Pike et al., 2002; Valentine et al., 2003; Wagstaff, MacVeigh, Scott, Brunas-Wagstaff, & Cole, 2003; Mecklenburg, 2006; for review, see Fawcett, Russell, Peace, & Christie, 2011), 87% of eyewitness experts agreed with the statement that "The presence of a weapon impairs an eyewitness's ability to accurately identify the perpetrator's face" and 77% would be willing to testify to that effect in court (Kassin, Tubb, Hosch, & Memon, 2001). Is weapon focus as strong and reliable as many eyewitness experts believe it to be?

An early meta-analysis by Steblay (1992) found that the presence of a weapon during a crime significantly reduced the correct ID rate, but the size of the effect was small. Nearly two decades later, a second meta-analysis found that the presence of a weapon during a crime caused a small, but significant, reduction in identification accuracy (Fawcett et al., 2011). Identification accuracy, in this sense, did not distinguish between changes in correct and false IDs, but combined correct IDs, correct rejections, false IDs and misses into an overall accuracy score. Although the meta-analyses have found a small effect, it is not clear whether the presence of a weapon during a crime impacts memory for the perpetrator. This is because many studies investigating weapon focus only measured changes in the correct ID rate (e.g. Loftus et al., 1987; Kramer et al., 1990; Shaw & Skolnick, 1994; Pickel, 1998; 1999) or only measured changes in the false ID rate (e.g. Maas & Kohnken, 1989; Hulse & Memon, 2006). A change in the correct ID rate (or the false ID rate) due to the presence of a weapon may reflect either a change in discriminability or a change in response bias (Wixted & Mickes, 2012). In order to determine whether weapon focus reflects a reduction in discriminability or a conservative shift in response bias, the correct *and* false ID rates must be measured. With correct and

false ID rates, receiver operating characteristic (ROC) analysis can then be conducted, which can determine if weapon focus impacts discriminability or response bias (Wixted & Mickes, 2012).

Gaps in Research

Ten studies (Cutler & Penrod, 1988; Cutler et al., 1987a, b; Cutler et al., 1986; O'Rourke et al., 1989; Carlson & Carlson, 2012; 2014; Erickson et al., 2014; Carlson et al., 2017; Carlson, Dias, Weatherford, & Carlson, 2016) have investigated weapon focus in perpetrator-present and perpetrator-absent lineups. However, of these ten studies, five did not report correct *and* false ID rates for weapon present and weapon absent conditions (Cutler & Penrod, 1988; Cutler et al., 1987a, b; Cutler et al., 1986; O'Rourke et al., 1989). Two studies (Carlson & Carlson, 2012; Erickson et al., 2014) reported correct and false ID rates for weapon present and weapon absent conditions, but did not conduct ROC analysis. Still, a clear reduction in discriminability was evident in both of these studies as the overall correct ID rate was significantly lower in the weapon present condition than the weapon absent condition, while the overall false ID rate was significantly higher in the weapon present condition than the weapon absent condition. There are only three weapon focus studies that have measured discriminability using ROC analysis (Carlson & Carlson, 2014; Carlson et al., 2016; Carlson et al., 2017). The presence of a weapon impacted discriminability (i.e. producing a lower ROC) in two of these studies (Carlson & Carlson, 2014; Carlson et al., 2016). The effect was not found in Carlson et al. (2017) and, as we shall see, Carlson and Carlson's (2012; 2014) findings are confounded.

Perceptual Analysis

I conducted a fine-grained perceptual analysis of the weapon present and weapon absent videos used in Carlson and Carlson (2012; 2014). I analysed frames that contained the perpetrator's face. There were 44 frames in the weapon absent video and 69 frames in the weapon present video that contained the perpetrator's face. Photo editing software (i.e. Adobe Photoshop CS) calculated the number of pixels the face occupied in each frame and the level of brightness for each face.

Size of the Face

Because the size of the face, as measured in pixels, was dependent on the size of the image, a percentage was calculated by taking the size of the face and dividing that by the size of the image. For example, if the face in a particular frame consisted of 10,000 pixels and the image was 640 x 360 pixels or, in other words, 230,400 pixels in total, then the face occupied 4.3% of the frame (i.e. 10,000/230,400 = 4.3%). The perpetrator's face in the weapon present video occupied, on average, 3.8% (*SD* = 1.8%) of each frame, whereas the perpetrator's face in the weapon absent video occupied, on average, 22.6% (*SD* = 7.8%) of each frame. That difference is highly significant, $t(111) = 19.65$, $p < 001$. The size of the face in the weapon present video was smaller because the perpetrator was filmed at a greater distance from the camera.

Brightness of the Face

The brightness of the perpetrator's face was measured for each frame and was averaged across frames to determine the overall brightness of the perpetrator's face in the weapon present and weapon absent videos. Brightness ranged from 0 (completely black) to 255 (completely white). Each pixel has a colour that falls somewhere between 0 and 255. Therefore, if the face was recorded in dim lighting, then the pixels that comprise the face will, on average, have a lower value (i.e. a value closer to 0) than if the face was recorded in bright lighting. The perpetrator's face in the weapon present video was darker ($M = 119.6$, $SD = 32.9$) than the perpetrator's face in the weapon absent video ($M = 153.9$, $SD = 32.4$) and that difference is highly significant, $t(111) = 5.46$, $p < .001$. The perpetrator's face in the weapon present video was darker because the perpetrator was standing further away from the hallway light.

Considering these differences, it is not surprising to observe greater discriminability in the weapon absent condition than the weapon present condition. It is reasonable to expect eyewitnesses to have better memory for the perpetrator when the perpetrator's face is closer to the camera and is shown in bright instead of dim light.

This perceptual analysis of the weapon present and weapon absent videos confounds Carlson and Carlson's (2012; 2014) findings. If these findings are ignored, then there are only two studies that have shown reduced discriminability for the perpetrator when the perpetrator previously held a weapon (Erickson et al., 2014; Carlson et al., 2016), which is hardly strong evidence of a weapon focus effect. In fact, there is a recent study from the same laboratory that found no difference in discriminability between weapon present and weapon absent conditions (Carlson et al., 2017). So, the question remains, does weapon focus impact discriminability or response bias?

Is Weapon Focus Meaningful to Judges and Jurors?

If weapon focus is reliably found to reduce discriminability, is this effect meaningful to judges and jurors? Intuitively, it may seem that informing judges and jurors of effects that reduce discriminability would improve jury decision making, but this is not necessarily the case. Judges and jurors should not be concerned with whether an effect reduces discriminability and, instead, should be concerned with the reliability of suspect IDs admitted as evidence in court. Because ROC analysis does not measure the reliability of a suspect ID, a more appropriate analysis for judges and jurors is confidence-accuracy characteristic (CAC) analysis (Mickes, 2015). CAC analysis plots suspect ID accuracy across varying levels of confidence (from low-confident identifications to high-confident identifications). In many cases, people can appreciate conditions that affect their memory and adjust their confidence accordingly (e.g., if they only saw the perpetrator for a short duration, they tend to be less likely to give high confidence identifications, but those high confidence identifications tend to be highly accurate). An eyewitness may or may not appreciate the fact that the presence of a weapon can impact memory accuracy. If eyewitnesses can appreciate that their memory is less accurate because of the presence of the weapon, then they should be able to use their confidence judgments more appropriately. This information can potentially be very useful to judges and jurors attempting to determine the veracity of a suspect ID.

Experiment 6

Although weapon focus has been studied extensively for several decades (e.g. Loftus et al., 1987), it is still unclear how weapon focus impacts lineup identification performance in terms of discriminability and reliability. The current experiment sought to determine whether weapon focus reflects a reduction in discriminability or a shift in response bias by conducting ROC analysis. In addition, CAC analysis was conducted to assess reliability.

Methods

Participants

Participants (N = 644) from the University of California, San Diego (UCSD) completed the experiment in exchange for course credit. Participants who failed to answer the validation question correctly (stating the crime committed; n = *75*) were excluded from all analyses. The remaining (n = 569; 132 male, 432 female, 5 unspecified; age in years: M = 20.4, SD = 2.5) participants watched a weapon present video *and* a weapon absent video and, thus, made two separate identification decisions from two separate lineups. Note that a within-participants design was chosen because it would help reduce any between-participants' error that may hide a weapon focus effect. For each lineup, participants were randomly assigned to be tested on either a perpetrator-present lineup (n = 562) or a perpetrator-absent lineup (n = 576). This means that, in total, we collected 1,138 identification decisions. The UCSD Institutional Review Board approved this study.

Materials

The study stimuli were third-person point of view video recordings of eight criminal scenarios (i.e. eight videos of different crimes each recorded from a third person perspective) that involved a unique perpetrator and witness for each video. The criminal situations included: two house burglaries, two carjackings, two muggings, and two public confrontations. The perpetrators were white men in their early twenties. The perpetrators' faces were shown for 6 – 9 seconds. Two versions for each criminal situation were recorded that included or did not include a weapon. Filmmakers did their best to record

the same scene with and without a weapon. In the weapon-absent conditions, the perpetrators clenched their fist, whereas, in the weapon-present conditions, the perpetrators either held a gun (a black, handheld pistol) or a knife. Filler selections later used in the lineups were based on the general similarities of the perpetrators' appearance, age, ethnicity, height, and weight. Fillers were collected from the Florida Correctional database (http://www.dc.state.fl.us/) from the supervised population information list.

Procedure

Following consent, participants were randomly assigned to a weapon present or weapon absent condition and watched a recording of one of the eight criminal scenarios. Participants then engaged in a 5-minute distractor task (a game of Tetris) before attempting to identify the perpetrator from a lineup. Participants were randomly assigned to a perpetrator-present or perpetrator-absent lineup that presented six photos simultaneously arranged in a 2 X 3 matrix. The position of the perpetrator in the perpetrator-present lineups was randomly determined for each participant and participants were instructed that the perpetrator may or may not be present in the lineup. No fillers were designated as the innocent suspect. Confidence in the identification decision was collected using an 11-point scale ranging from 0 (just guessing) to 100 (absolutely certain). Following their identification decision, participants answered several questions about the crime, including a validation question (what crime was committed?), and were asked to supply demographic information.

Participants then watched another crime video that either included or did not include a weapon (i.e. every participant viewed two different crime videos, one of which included a weapon). Following the crime video, participants took part in the same distractor task for 5 minutes and attempted to identify the perpetrator from a simultaneous lineup. After making an identification decision, participants answered general questions about the crime video (which included a validation question) and were then fully debriefed.

Results

Descriptive Analysis

Those who did not answer the validation questions correctly were excluded from all analyses (n = 75 participants). In Table 1, the type of weapon used is displayed as well as the *d′* values for the weapon-present and weapon-absent conditions in each video. A perceptual analysis was then conducted.

Table 1

The weapon, *d′* values, and the difference between those values displayed for each video.

Video	Weapon	W Present	W Absent	Difference	WF Trend
Video 1	Knife	1.42	1.89	-.47	Yes
Video 2	Gun	.92	1.00	-.08	Yes
Video 3	Gun	2.22	2.03	.19	No
Video 4	Knife	.26	.63	-.37	Yes
Video 5	Knife	1.03	1.63	-.60	Yes
Video 6	Knife	2.31	2.12	.19	No
Video 7	Gun	2.04	1.84	.15	No
Video 8	Knife	1.98	2.38	-.40	Yes

Note: W Present, weapon present; W Absent, weapon absent. The weapon focus trend was measured by subtracting the *d′* for the weapon present condition by *d′* for the weapon absent condition. These differences were not significant.

Perceptual Analysis

A perceptual analysis was conducted of the eight weapon present videos and the eight weapon absent videos used in this experiment. This analysis sought to determine whether the size and the brightness of a perpetrator's face was significantly different between the weapon present and weapon absent videos. Adobe Photoshop CS was used to conduct this analysis. Every frame from the eight crime videos that contained a face of a perpetrator was used in this analysis. In total, there were 1,516 frames that contained the perpetrator's face across the weapon present videos and 1,268 frames that contained the

perpetrator's face across the weapon absent videos. Note that a perceptual analysis of each crime video is presented in the Appendix. This includes a perceptual analysis of the size of the perpetrator's face for each video and a perceptual analysis of the brightness of the perpetrator's face for each video.

The size of the perpetrator's face was measured in pixels and that number was divided by the total number of pixels in the frame. The perpetrator's face occupied, on average, 21.6% (*SD* = .17%) of the frame across the eight weapon present videos, whereas the perpetrator's face occupied, on average, 22% (*SD* = .15%) of the frame across the eight weapon absent videos. That difference was not statistically significant, $t(2782) = -.621$, $p = .535$, which means that the size of the perpetrator's face across the eight crime videos was virtually the same between the weapon present condition and the weapon absent condition.

The brightness of the perpetrator's face was measured for each frame from the eight weapon present videos and the eight weapon absent videos. The perpetrator's face across the eight weapon present videos was slightly brighter ($M = 112.1$, $SD = 42.0$) than the perpetrator's face across the eight weapon absent videos ($M = 110.2$, $SD = 40.4$), but that difference was not significant, $t(2782) = .818$, $p = .413$. Thus, the perpetrator's face across the weapon present videos was just as bright as the perpetrator's face across the weapon absent videos.

Because the perceptual analysis revealed no significant difference in the size and the brightness of the perpetrator's face across the eight crime videos, the data from each video were collapsed into an overall weapon present and weapon absent condition. Response frequencies for false IDs, filler IDs, misses, correct rejections, and correct IDs for each level of confidence for weapon present and weapon absent conditions are displayed in Table 2.

Table 2

Response frequencies for every decision outcome displayed for each level of confidence.

	Weapon Present					Weapon Absent				
	Perp Present			Perp Absent		Perp Present			Perp Absent	
Confidence	CID	FID	noID	FIDs	noID	CID	FID	noID	FID	noID
0	0	3	3	5	5	0	5	2	1	6
10	0	1	2	6	1	3	2	1	2	2
20	2	5	3	8	1	5	1	1	5	2
30	4	12	0	18	3	5	7	4	8	7
40	8	5	6	14	8	10	5	4	10	14
50	11	13	10	19	14	16	9	9	23	19
60	17	9	7	21	19	19	8	10	20	19
70	23	6	10	29	22	27	8	12	16	24
80	24	5	7	17	20	33	3	11	20	24
90	29	2	5	13	23	21	2	10	14	18
100	30	2	5	5	18	26	2	7	3	10

Note: Perp, perpetrator; ID, identification; CIDs, correct IDs; FIDs, filler IDs.

ROC Analysis

The suspect ID rates for perpetrator-present lineups (i.e. correct ID rates), suspect ID rates for perpetrator-absent lineups (i.e. false ID rates), and the filler ID rates for both perpetrator-present and perpetrator-absent lineups for each level of confidence are shown in Table 3. The italicized values were used to construct the ROC curves shown in Figure 1. Because there was no designated innocent suspect, the false ID rate was estimated by dividing by 6. This is because there were 6 fillers in the perpetrator-absent lineup (see Chapter 2).

Table 3

Suspect IDs, filler IDs and no IDs for perpetrator-present and perpetrator-absent lineup rates by level of confidence per condition.

		Weapon Present			Weapon Absent		
	Confidence	SID	FID	noID	SID	FID	noID
Perp Present	0	***.55***	.23		***.50***	.18	
	10	*.55*	.22		*.49*	.16	
	20	*.55*	.22		*.47*	.16	
	30	*.54*	.20		*.45*	.15	
	40	*.53*	.16		*.40*	.13	
	50	*.50*	.14	.22	*.33*	.11	.29
	60	*.46*	.09		*.25*	.08	
	70	*.39*	.06		*.16*	.05	
	80	*.31*	.03		*.09*	.02	
	90	*.22*	.02		*.04*	.01	
	100	*.11*	.01		*.01*	.01	
Perp Absent	0	***.08***	.50		***.08***	.46	
	10	*.08*	.49		*.08*	.45	
	20	*.08*	.47		*.07*	.45	
	30	*.07*	.44		*.07*	.43	
	40	*.06*	.38		*.07*	.40	
	50	*.06*	.34	.53	*.06*	.36	.48
	60	*.05*	.28		*.05*	.27	
	70	*.04*	.21		*.03*	.20	
	80	*.02*	.11		*.02*	.14	
	90	*.01*	.06		*.01*	.06	
	100	*.00*	.02		*.00*	.01	

Note: Perp, perpetrator; ID, identification; SIDs, suspect IDs; FIDs, filler IDs.

Figure 1 shows the ROC curves for the weapon present and weapon absent conditions. In order to calculate the *pAUCs*, the most conservative false alarm rate from the weapon absent condition (.08) was used because both ROC curves extend to this cutoff (Robin et al., 2011). Selecting the most liberal false alarm rate obtained from the weapon-present condition would require a portion of the ROC for the weapon absent condition to be extrapolated to that cutoff. The *pAUC* for the weapon present condition (.038) was not significantly different than the *pAUC* for the weapon absent condition (0.029), $D = 0.55$, $p = .60$. Though there was no significant difference between the two curves, the weapon present curve does look higher at higher levels of confidence (i.e., more conservative responding). Because of this, we compared the *pAUCs* of both conditions using a smaller false ID cutoff (.035). Using this cutoff, the *pAUC* for the weapon present condition (.008) was not significantly different than the *pAUC* for the weapon absent condition (.007), $D = 0.05$, $p = .97$.

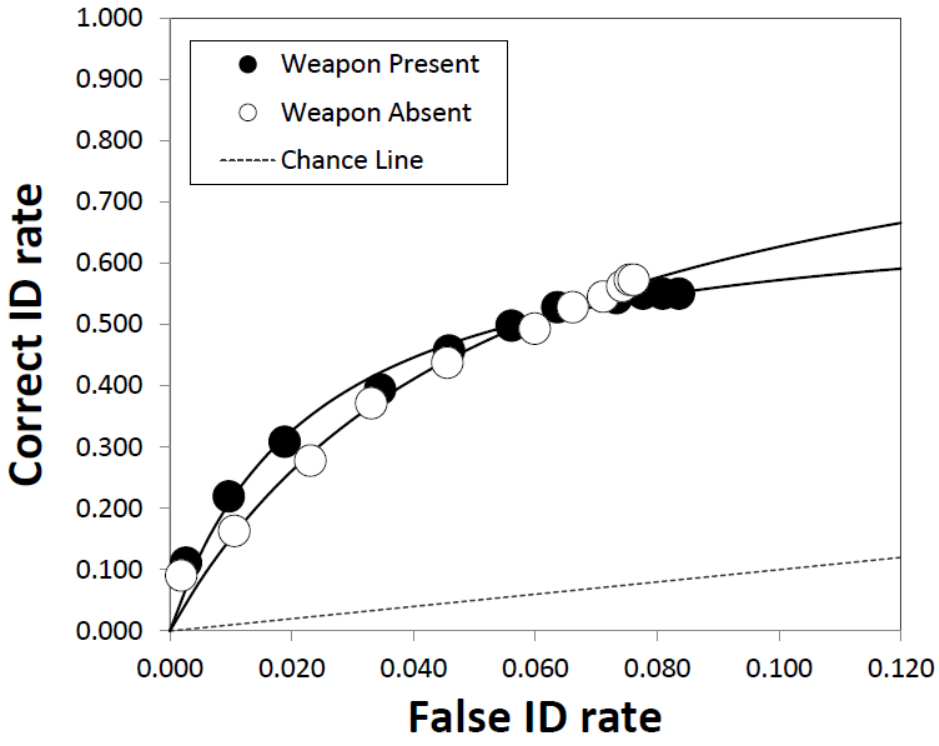

Figure 1. Receiver operating characteristic (ROC) curves for the Weapon Present and Weapon Absent conditions. The lines through the ROC curves were estimated using a hyperbolic function. The dashed line represents chance performance.

CAC Analysis

The relationship between confidence and accuracy was examined using CAC analysis (Mickes, 2015). The CAC plots are shown in Figure 2. Confidence was binned into low (0-60), medium (70-80), and high (90-100) confidence because there were too few identifications at some levels. Participants who expressed medium or high confidence in selecting the suspect from the weapon present and weapon absent condition seem to calibrate memory accuracy of the perpetrator with confidence in their suspect identification decision equally well. Whereas, those participants in the weapon present condition who expressed low confidence may have been less accurate than those in the weapon absent condition. The data appear too noisy at low levels of confidence to make any firm conclusions.

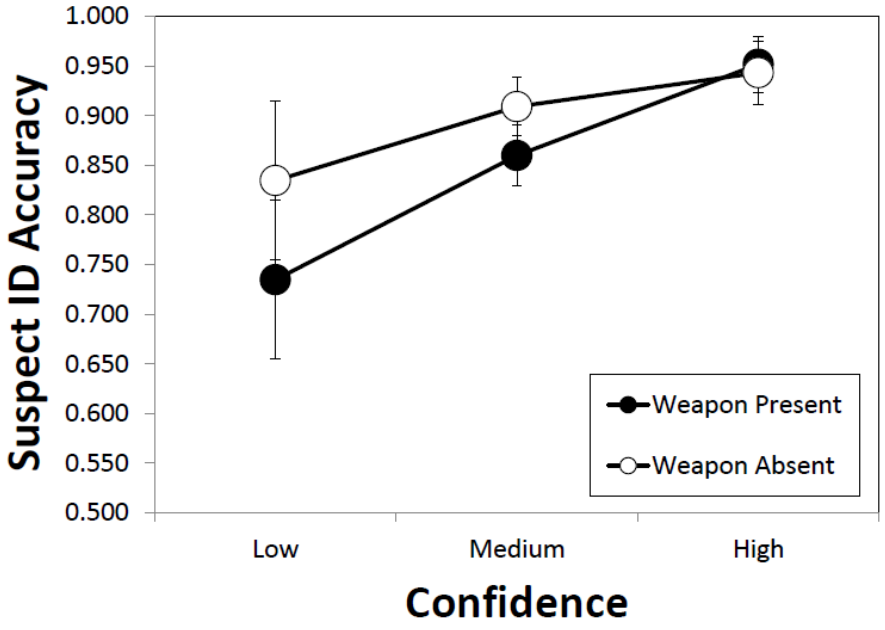

Figure 2. Confidence-accuracy characteristic (CAC) curves for the Weapon Present and Weapon Absent conditions. The bars represent standard error bars.

Discussion

The presence of a weapon has been found to adversely impact an eyewitness's ability to correctly identify the perpetrator from a lineup (e.g. Loftus et al., 1987). Although this effect is often not found (e.g., Cutler & Penrod, 1988; Cutler et al., 1986),

two meta-analyses have found a small, significant effect (Steblay, 1992; Fawcett et al., 2011). Yet, many of the studies that have found an effect, have only reported differences in the correct ID rate between weapon present and weapon absent conditions. A difference in correct ID rates may reflect a reduction in discriminability or a conservative shift in responding (Wixted & Mickes, 2012). Despite this ambiguity in how a weapon might impact identification performance, eyewitness experts largely agree that the weapon focus effect is a reliable effect and many are willing to testify in court that a weapon causes an eyewitness to be less able to accurately identify the perpetrator from a lineup (Kassin et al., 2001).

In Experiment 6, we investigated the weapon focus effect in perpetrator-present and perpetrator-absent lineups in order to measure correct and false ID rates. With correct and false ID rates, ROC analysis can then be conducted and differences in discriminability and response bias are evident in ROC curves. Our results indicate that the presence of a weapon during a crime does not impact an eyewitness's ability to discriminate innocent from guilty suspects, which does not replicate the only other published weapon focus study in which ROC analysis was conducted (Carlson and Carlson, 2014). Although no significant difference was found, the ROC for the weapon present condition appeared to be slightly higher than the ROC for the weapon absent condition across high levels of confidence. A separate *pAUC* analysis was conducted, but revealed no significant differences in discriminability across high levels of confidence. Critically, it may have been the case that a weapon focus effect occurred for some stimuli, but the process of combining the data into an overall ROC curve hid an effect. Although none of the d' values in the weapon present condition differed significantly from the d' values in the weapon absent condition (see Table 1), a few videos show a lower d' in the weapon present condition and a higher d' in the weapon absent condition. Because there is not enough data to conduct *pAUC* analysis for each video, more data is needed to see if these trending effects are actually significant.

The CAC analysis shows that the presence of a weapon does not impact an eyewitness's ability to confidently and accurately identify the perpetrator from a lineup.

Despite the willingness of many eyewitness experts to testify in court on the deleterious effects of weapon focus (Kassin et al., 2001), these data suggest, based on CAC analysis, that the presence of a weapon does not impact the reliability of a suspect ID.

Experiment 7

Although an overall weapon focus effect (in terms of discriminability) was not found in Experiment 1, it could be the case that the process of combining the data across the varied set of stimuli masked a true weapon focus effect. Two videos that showed a trend towards a weapon focus effect (video 2 and video 5; see Table 5) were selected as the stimuli in this experiment. Will the weapon focus trend continue to hold once more participants are recruited?

Methods

Participants

Participants (N = 630) from the University of California, San Diego completed the experiment in exchange for course credit. Participants that failed to answer the validation question correctly (stating the crime committed; $n = 46$) were excluded from all analyses. The remaining (n = 584; 157 male, 420 female, 7 unspecified; age in years: M = 20.3, SD = 2.5) participants watched a weapon present video *and* a weapon absent video and, thus, made two separate identification decisions from two separate lineups. However, for each lineup, participants were randomly assigned to be tested on either a perpetrator-present lineup (n = 578) or a perpetrator-absent lineup (n = 590). This means that, in total, we collected 1,168 identification decisions. The UCSD Institutional Review Board approved this study.

Materials

The study stimuli were third-person point of view video recordings of two criminal scenarios (i.e. a mugging and a public confrontation) that were originally presented in Experiment 1 and found to show a trend toward a weapon focus effect.

Procedure

The procedure was the same as in Experiment 6.

Results

Descriptive Analysis

Those who did not answer every validation question correctly were excluded from all analyses (n = 46 participants). Response frequencies for false IDs, filler IDs, misses, correct rejections, and correct IDs for each level of confidence for weapon present and weapon absent conditions are displayed in Table 4. In Table 5, *d′* values are displayed for weapon present conditions and weapon absent conditions for both videos.

Table 4

Response frequencies for every identification decision outcome are displayed for each level of confidence.

	Weapon Present					Weapon Absent				
	Perp Present			Perp Absent		Perp Present			Perp Absent	
Confidence	CID	FID	noID	FID	noID	CID	FID	noID	FID	noID
0	0	5	4	6	4	1	3	2	3	3
10	4	4	0	7	4	2	5	1	7	0
20	4	5	2	5	6	2	1	3	7	2
30	7	16	6	14	7	4	8	1	24	12
40	6	12	5	25	16	11	10	2	14	4
50	17	11	8	32	22	12	11	11	20	15
60	11	12	8	24	11	18	15	9	26	19
70	25	17	7	16	21	18	15	15	23	34
80	18	5	10	7	16	17	4	10	6	16
90	17	4	4	4	23	20	7	9	8	18
100	24	3	8	5	20	27	1	6	5	23

Note: Perp, perpetrator; ID, identification; CIDs, correct IDs; FIDs, filler IDs.

Table 5

The weapon, *d'* values in the weapon present and weapon absent condition, and the difference between those values are displayed for both videos.

Video	Weapon	W Present	W Absent	Difference	WF Trend
Video 2	Gun	1.01	.95	.06	No
Video 5	Knife	1.60	1.60	.00	No

Note: The weapon focus trend was measured by subtracting the *d'* value for the weapon present condition by the *d'* value for the weapon absent condition.

ROC Analysis

The suspect ID rates for perpetrator-present lineups (i.e. correct ID rates), suspect ID rates for perpetrator-absent lineups (i.e. false ID rates), and the filler ID rates for both perpetrator-present and perpetrator-absent lineups for each level of confidence are shown in Table 6. The italicized values were used to construct the ROC curves shown in Figure 3. Because there was no designated innocent suspect, the false ID rate was estimated by dividing by 6. This is because there were 6 fillers in the perpetrator-absent lineup (see Chapter 2).

Table 6

Suspect IDs, filler IDs and no IDs for perpetrator-present and perpetrator-absent lineup rates by level of confidence per condition.

		Weapon Present			Weapon Absent		
	Confidence	SID	FID	noID	SID	FID	noID
Perp Present	0	***.46***	.33		***.47***	.28	
	10	*.46*	.31		*.47*	.27	
	20	*.44*	.29		*.46*	.26	
	30	*.43*	.28		*.45*	.25	
	40	*.41*	.22		*.44*	.22	
	50	*.39*	.18	.21	*.40*	.19	.25
	60	*.33*	.14		*.36*	.15	
	70	*.29*	.10		*.29*	.10	
	80	*.20*	.04		*.23*	.04	
	90	*.14*	.02		*.17*	.03	
	100	*.08*	.01		*.10*	.00	
Perp Absent	0	***.08***	.50		***.08***	.50	
	10	*.08*	.47		*.08*	.49	
	20	*.08*	.45		*.08*	.46	
	30	*.07*	.43		*.07*	.44	
	40	*.06*	.38		*.06*	.35	
	50	*.05*	.30	.52	*.05*	.30	.52
	60	*.03*	.19		*.04*	.24	
	70	*.02*	.11		*.02*	.15	
	80	*.01*	.05		*.01*	.07	
	90	*.01*	.03		*.01*	.05	
	100	*.00*	.02		*.00*	.02	

Note: Perp, perpetrator; ID, identification; SIDs, suspect IDs; FIDs, filler IDs.

Discriminability was measured by conducting ROC analysis and comparing the *pAUCs* as was previously done in Experiment 1. Figure 3 shows the ROC curves for the weapon present and weapon absent conditions collapsed across both videos. The *pAUC* for the weapon present condition (0.008) was not significantly different than the *pAUC* for the weapon-absent condition (0.008), $D = 0.009$, $p = .99$. This replicates the findings from the previous experiment: no weapon focus effect was found.

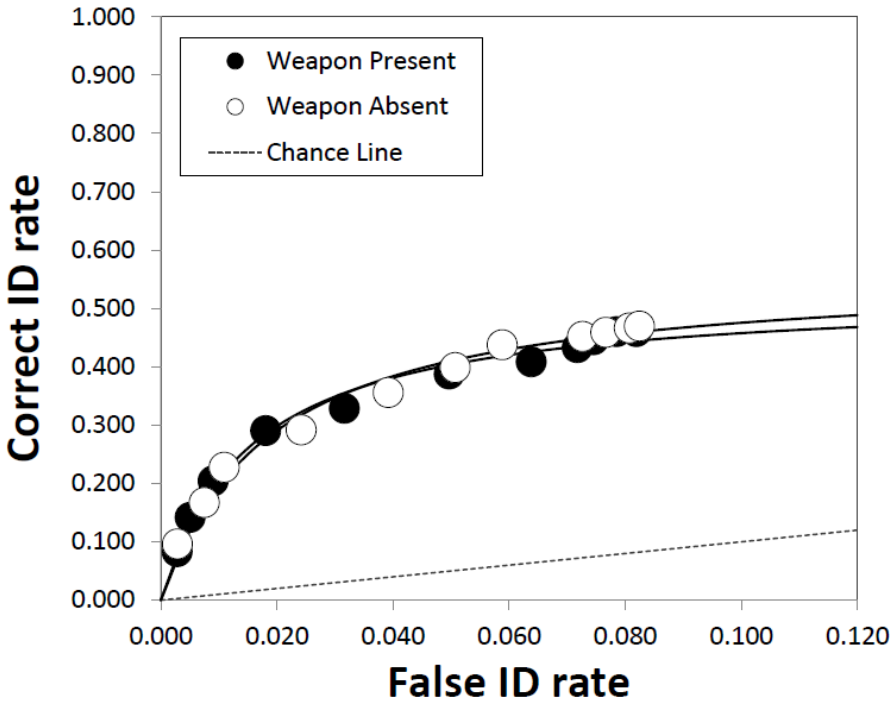

Figure 3. Receiver operating characteristic (ROC) curves for the Weapon Present and Weapon Absent conditions. The lines through the ROC curves were estimated using a hyperbolic function. The dashed line represents chance performance.

CAC Analysis

The CAC plots are shown in Figure 4. Confidence was binned into low (0-60%), medium (70-80%), and high (90-100) confidence. For each level of confidence, suspect ID accuracy (A) = # correct suspect IDs / (# correct suspect IDs + # innocent suspect IDs). Participants who expressed low, medium, or high confidence in selecting the suspect from the weapon present and weapon absent condition calibrated their memory accuracy of the

perpetrator with confidence in their suspect identification decision equally well. This replicates the CAC results from Experiment 6.

Figure 4. Confidence-accuracy characteristic (CAC) curves for the Weapon Present and Weapon Absent conditions. The bars represent standard error bars.

General Discussion

Because a large variety of videos were used as stimuli in Experiment 6 and data from each video were subsequently collapsed into an overall ROC and CAC, it is possible that a weapon focus effect occurred for some of the videos but such an effect was hidden due to collapsing the data across the wide variety of videos. In Experiment 7, two videos were selected from Experiment 1 that showed non-significant differences in d' that trended towards a weapon focus effect (i.e. video 2 and video 5; see Table 1). After more data was collected, the differences in d' were reduced and ROC analysis revealed no significant difference in discriminability between weapon present and weapon absent conditions for each video.

The results from Experiment 7 replicate the results from Experiment 6; no weapon focus effect was found. Although two meta-analyses (Steblay, 1992; Fawcett et al., 2011) of the weapon focus literature have been conducted, it is still not clear whether the presence of a weapon impacts discriminability. Carlson et al.'s (2017) findings and the findings from these two experiments suggest that a weapon (whether a gun or a knife) has no effect on discriminability, but a previous experiment conducted by Carlson and colleagues (Carlson et al., 2016) has shown a weapon focus effect (i.e. reduced discriminability in the weapon present condition). This issue might be resolved by conducting a registered replication report that attempts to replicate Loftus et al.'s (1987) original findings (similar to how the first registered replication report attempted to replicate the verbal overshadowing effect from Schooler and Engstler-Schooler, 1990). However, a perpetrator-absent lineup should be included in order to measure correct and false ID rates, and ROC analysis should be conducted. This way, the registered replication report could determine, across many independent laboratories, whether the presence of a weapon causes a reduction in discriminability.

Regardless of whether the presence of a weapon was found to reduce discriminability, triers of fact ought to be interested in whether the presence of a weapon impacts the reliability of a suspect ID. At the moment, eyewitness experts are largely willing to testify that the presence of a weapon impairs the ability of the eyewitness to accurately identify the perpetrator from a lineup (Kassin et al., 2001). However, CAC analysis conducted in Experiments 6 and 7 indicate that eyewitnesses who were exposed to a weapon were just as able to confidently and accurately identify the perpetrator as those who were not exposed to a weapon. This means that eyewitnesses who make a highly confident suspect ID after being exposed to a weapon are reliable and that is important for judges and jurors to know.

Chapter 6

Confidence is the Best Available Marker of Suspect ID Accuracy

In criminal trials it is often the case that eyewitness identification evidence is the strongest or only evidence available to prosecutors (Goldstein et al., 1989). Triers of fact (i.e. judges and jurors) tend to place a great deal of faith in this evidence in determining the guilt or innocence of a suspect (Loftus, 1975, 1979; Ellison & Buckhout, 1981; Lindsay, Wells, & Rumpel, 1981). This can be problematic because eyewitnesses have been known to make mistakes despite honest attempts to correctly identify the perpetrator. For example, the Innocence Project, whose mission it is to lobby on behalf of wrongfully convicted individuals, has helped release 334 innocent suspects through the use of DNA testing and have found false identifications (false IDs) to be the leading cause of these wrongful convictions (Innocence Project, 2017). With this problem in mind, eyewitness identification researchers have explored several possible markers of eyewitness identification accuracy that may be able to reliably differentiate between a correct ID and a false ID (e.g. Sporer, 1993; Smith, Lindsay, & Pyrke, 2000).

Confidence in the Identification Decision

The confidence an eyewitness expresses in an identification decision happens to be the most extensively researched as well as the most controversial marker of eyewitness identification accuracy (Wells & Murray, 1983; Lindsay, 1986; Sporer, Penrod, Read, & Cutler, 1995; Mickes, 2015). Much of the early research has shown that confidence can easily become inflated irrespective of identification accuracy. For instance, an eyewitness may increase their confidence in their identification decision after hearing that other eyewitnesses have identified the same suspect (Luus & Wells, 1994) or by being exposed to the same suspect again (Brown, Deffenbacher & Sturgill, 1977). An early review of the literature found a weak relationship between confidence and accuracy and determined that confidence is “functionally useless in forensically representative settings” (p. 165, Wells & Murray, 1984).

A more recent meta-analysis found the confidence-accuracy (CA) relationship to be noticeably stronger when the analysis was limited to eyewitnesses who chose a lineup member (i.e. excluding those that did not select a lineup member) and when confidence was measured at the time of the initial identification (Sporer et al., 1995). Limiting the analysis to "choosers" is reasonable because only choosers end up testifying in court. Measuring confidence at the time of the initial identification is reasonable as well because there is less of an opportunity for confidence to become inflated at that time (Wixted et al., 2015). Still, Sporer et al. (1995) emphasized that "… confidence is far from a perfect indicator of witness accuracy" (p. 324).

The meta-analyses by Wells and Murray (1984) and Sporer et al. (1995) have interpreted the CA relationship as weak and modest, respectively, based on low values derived from the point-biserial correlation coefficient. Juslin et al. (1996) point out, though, that this value is flawed. A low point-biserial correlation does not necessarily mean a weak CA relationship. The point-biserial correlation coefficient is a useful effect-size statistic for measuring how well accuracy (correct vs. incorrect IDs) predicts eyewitness confidence, but triers of fact are interested in the opposite question of how well eyewitness confidence predicts accuracy. This means that the point-biserial correlation coefficient can fluctuate on a wide range of values even when confidence is perfectly calibrated with accuracy. Juslin et al. (1996) have analysed the CA relationship using calibration curves. Analysing the CA relationship in this way shows a strong relationship (e.g. Brewer & Wells, 2006; Brewer & Palmer, 2010).

Yet, calibration analysis is limited in its use to triers of fact. This is because calibration analysis includes filler IDs in measurements of accuracy. Fillers are known innocent members presented alongside the police suspect in a lineup; they are not suspects. CAC analysis (Mickes, 2015), on the other hand, excludes filler IDs, showing the relationship between confidence and accuracy for identified suspects. It is this information that is most informative to triers of fact. When CAC analysis is conducted, highly confident eyewitnesses in the laboratory (e.g. Mickes, 2015) and in the field (Wixted et al., 2016) are shown to be highly accurate. Thus, although early reviews of the scientific

literature have found confidence to be a generally poor marker of identification accuracy (e.g. Wells & Murray, 1983), recent analyses convincingly show that confidence, when properly measured, is actually a strong predictor of identification accuracy (Mickes, 2015).

Visual Behaviour

The belief that confidence is uninformative of accuracy has spurred researchers to investigate other possible markers of eyewitness identification accuracy. So far, several markers have been identified as being somewhat capable of differentiating between correct and false IDs. These include the accuracy of the description of the perpetrator (e.g. Pigott & Brigham, 1985), the time it takes to identify a lineup member (e.g. Sporer, 1992; 1993) and the visual behaviour used to examine the faces in the lineup (e.g. Flowe & Cotrell, 2011). Of these three markers, analysing visual behaviour while scanning a face seems particularly promising because eye movements are an essential component of studying and recognizing a face.

Visual Behaviour during the Study Phase

Henderson, Williams, and Falk (2005) demonstrated this by recording eye movements while participants studied a list of faces. Participants were instructed to either fixate at the centre of each face, thereby restricting their eye movements, or were allowed to freely view each face. During the recognition test, studied "old" faces (i.e. targets) and novel "new" faces (i.e. lures) were randomly presented one at a time for an old or new recognition decision. A correct response for a target is old; these responses are called "hits." A correct response for a lure is new; these responses are called "correct rejections." Incorrectly declaring a target as new is a "miss" and incorrectly declaring a lure as old is a "false alarm" (See Chapter 1 for review). Discriminability was significantly lower when eye movements were previously restricted. Henderson et al. suggested that by restricting eye movements, participants were less able to encode the crucial facial information contained outside of foveal vision (i.e. outside the sharp, but central vision of the eye) and had worse recognition memory as a result.

Visual Behaviour during the Recognition Phase

Differences in visual behaviour are also seen for old and new faces during recognition tests. For instance, within the first five seconds of viewing a face, participants fixate more on a new face than on an old face, but the distance between these fixations is smaller (i.e. their eye movements are more constrained) and participants end up sampling more regions of a new face (i.e. regions closer to the eyes, nose, and mouth; Althoff & Cohen, 1999). Because of the stark differences in visual behaviour when participants study and attempt to recognize faces, it is possible that there are differences in visual behaviour when an eyewitness makes a correct or a false ID. Perhaps, suboptimal scanning of an innocent suspect during a lineup causes participants to falsely recognize that face. Analysing eyewitnesses' visual behaviour may therefore help distinguish between a correct and false ID.

Visual Behaviour in Eyewitness Studies

Forensic studies have utilized eye tracking methods to shed light on the memorial processes used when witnessing a crime (e.g. Loftus et al., 1987) and, more recently, when identifying a suspect from a lineup (e.g. Mansour, Lindsay, Brewer, & Munhall, 2009; Flowe, 2011; Flowe & Cotrell, 2011). Flowe and Cottrell (2011) had participants study a list of computer generated faces (i.e. targets). After a brief delay, participants attempted to identify the targets from a series of target-present and target-absent simultaneous lineups. Their eye movements were recorded for each lineup.

There were several differences in visual behaviour when participants made correct and false IDs. First, falsely identified innocent suspects received more fixations during the lineup, on average, than correctly identified perpetrators. Second, upon first viewing the faces in the lineup (i.e. the first fixations made to each face in the lineup), more time was spent fixating on identified perpetrators than identified innocent suspects. However, once participants began to revisit previously scanned faces, more time went back to fixating on the identified innocent suspect than the identified perpetrator. Third, the length and number of return visits to the identified innocent suspect indicate that participants took more time, in total, to identify the innocent suspect than to identify the perpetrator. This

replicates previous studies which have found that perpetrators are identified faster than innocent suspects (e.g. Sporer, 1992; 1993). Perhaps participants felt more certain when identifying the perpetrator than when identifying the innocent suspect and thus did not the feel the need to spend more time looking back and forth between similar looking faces to make an identification.

In a similar study, Mansour et al. (2009) had participants watch twelve short mock crime videos, each depicting a unique perpetrator committing the same crime. After a brief delay, participants then attempted to identify the perpetrators from a series of perpetrator-present and perpetrator-absent simultaneous lineups. Eye movements were recorded during the identification process. Mansour et al. compared the number of first-order and second-order comparisons made for correct and false IDs. A first-order comparison occurs when participants fixate on a face (A) and then fixate on another face (B) before fixating back on the previous face (A). A second-order comparison occurs when participants fixate on a face (A), then fixate on a second face (B), and then a third face (C), before fixating back onto the first face (A). Mansour et al. hypothesized that participants will make more first-order and second-order comparisons when making a false ID than when making a correct ID. This is because first-order and second-order comparisons suggest deliberation in the decision process and participants are expected to deliberate more when making a false ID. This hypothesis was partially supported. Correct IDs and false IDs received the same number of first-order comparisons, but false IDs received significantly more second-order comparisons than correct IDs.

Although the literature is small, consisting of only a handful of studies, recording participants' visual behaviour has shown some systematic differences between correct and false IDs (Flowe, 2011). Namely, participants fixate more, and for a longer duration, on innocent suspects who are falsely identified than they do for perpetrators who are correctly identified. Yet, there may be more information to glean from the eyes of eyewitnesses. It has recently been shown that pupil dilations in the eye could reflect memorial processes in standard recognition tasks (e.g. Vo, Jacobs, Kuchinke, Hofmann, Conrad, Schacht, & Hutzler, 2008). To the best of my knowledge, no eyewitness identification study has

analysed pupil dilations when participants witness a perpetrator committing a crime or when participants attempt to identify a perpetrator from a lineup. It is possible that pupil dilations and eye movements can both serve as robust markers of eyewitness identification accuracy.

Pupil Dilations

Early cognitive psychology research dating back to the 1960s has investigated slight changes in pupil size. Psychosensory fluctuations in pupil size are small (no more than 0.5 - 1.0mm) and are thus difficult to see with the naked eye (Beatty, 1982b). Yet, these slight but consistent pupillary changes appear to reflect cognitive processes occurring in the brain (Beatty, & Lucero-Wagoner, 2000). Described as "a permanently implanted electrode" and "the only visible part of the brain" (Janisse, 1977, p. 1), the pupil has been able to serve as a physiological indicator of psychological processes in a wide array of tasks (Beatty, & Lucero-Wagoner, 2000). Recently, researchers have sought to determine whether pupil dilations can serve as a physiological indicator of recognition memory processes in standard list-learning recognition tasks (e.g. Maw & Pomplun, 2004; Otero et al., 2006; Ryan, Hannula, & Cohen, 2007; Vo et al., 2008; Goldinger, He, & Papesh, 2009; Kafkas & Montaldi, 2011; Heaver & Hutton, 2011; Papesh & Goldinger, 2012).

Pupil Dilations and Recognition Memory

According to dual-process theorists, two processes contribute to a recognition decision: recollection and familiarity (e.g. Wixted & Mickes, 2010). Recollection consists of retrieving specific details associated to an item, whereas familiarity allows one to confidently appreciate the fact that an item or event was previously experienced even though no contextual detail can be retrieved (for review see Yonelinas, 2002). In a series of experiments, Otero, Weekes, and Hutton (2011) investigated whether pupils dilated to recollection-based or to familiarity-based responses. In Experiment 1, participants studied a list of words for an upcoming recognition test. During the recognition test, old and new words were randomly presented one at a time for an old or new recognition decision. For every item that was deemed old, participants would respond "remember" when

experiencing recollection and "know" when experiencing familiarity (for review of the remember/know procedure, see Tulving, 1985). If participants believed the item was not previously studied, then they would respond "new". Participants' pupils were measured when making these judgments.

If participants' pupil dilations reflect recognition memory processes (i.e. recollection and familiarity), then participants who respond "remember" should have greater pupil dilations than participants who respond "know". Remember judgments are thought to reflect stronger memories, on average, than know judgments as remember judgments are reliably associated with higher confidence, higher accuracy, and faster reaction times than know judgments (e.g. Dunn, 2004; Rotello & Zeng, 2008; Wixted & Stretch, 2004). More generally, old items should elicit greater pupil dilations than new items because old items have been previously studied. The results showed that old items elicited greater pupil dilations than new items, replicating previous studies that have found a similar pupil old/new effect (Gardner, Mo, & Borrego, 1974; Gardner, Mo, & Krinsky, 1974; Maw & Pomplun, 2004; Otero et al., 2006; Vo et al., 2008). In addition, "remember" responses to old items elicited greater pupil dilation than "know" responses to old items. "Remember" responses also happened to be significantly more accurate than "know" responses (i.e. participants were much more likely to respond "know" to a new item than to respond "remember" to a new item). Together, these results suggest that the extent to which the pupil dilates during a recognition task contains information about 1) how strongly the participant recognizes the item and 2) how accurate they are in their recognition decision.

In Experiment 2, Otero et al. manipulated memory strength for old items by having participants study a list of words in either deep encoding or shallow encoding conditions. In the shallow encoding condition, participants were instructed to count the number of syllables for each word, whereas in the deep encoding condition, participants were instructed to generate a synonym for each word. Similar procedures have produced strong memories for words that were encoded deeply and have produced weak memories for words that were encoded shallowly (e.g. Craik & Lockhart, 1972; Craik & Tulving, 1975).

After studying the list of items (either under deep or shallow encoding instructions), participants then took part in a recognition task where old and new items were randomly presented one at a time for an old or new recognition decision. Participants' pupils were recorded when making a recognition decision. As in Experiment 1, a pupil old/new effect was observed; pupil dilations were greater for old items than for new items. However, words that were encoded deeply (i.e. the stronger memories) elicited larger pupil dilations than words that were encoded shallowly (i.e. the weaker memories).

Experiments 1 and 2 have thus far demonstrated that 1) the pupil dilates more for old items than for new items, and 2) the extent to which the pupil dilates contains information about how strongly the participant recognizes the item and 3) how accurate they are in their recognition decision. Experiment 3 sought to determine whether the pupil dilates more for items that were correctly recognized than for items that were falsely recognized. In this experiment, participants studied lists of exemplars from several different categories. For example, participants studied items such as *bear*, *lion*, *giraffe*, *monkey*, *mouse*, and *camel* which are exemplars from the category *mammals*. During the recognition test, old exemplars (e.g. *bear*), new exemplars (e.g. *zebra*), and new unrelated items (e.g. *flower*) were randomly presented one at a time to participants for an old or new recognition decision. As in Experiments 1 and 2, a significant pupil old/new effect was observed; old items (e.g. *bear*) elicited greater pupil dilations than new items (e.g. *zebra* and *flower*). What is of interest though in this experiment is, after participants recognized an item as old, whether their pupils dilated more for items that were correctly recognized (e.g. correctly recognizing *bear* as old) than for items that were falsely recognized (e.g. falsely recognizing *zebra* as old). Pupil dilations were, in fact, greater for items that were correctly recognized than for items that were falsely recognized and this was true even when participants falsely recognized new exemplar items (e.g. *zebra*), which were highly similar to previously studied exemplars. Together, these results suggest that, when an eyewitness identifies a suspect from a lineup with a certain amount of confidence, the pupil will dilate more for the identified perpetrator than for the identified innocent suspect.

Pupil Dilations and Decision-Making

Many studies over the past several decades have shown that changes in pupil size can be linked to participants' choice during perceptual, cognitive, and economic decision-making tasks (e.g. Hess & Polt, 1964; Kahneman & Beatty, 1966; Simpson & Hale, 1969). This decision-related pupil dilation has been specifically linked to the formation of a final decision (Einhauser, Stout, Koch, & Carter, 2008; Einhauser, Koch, & Carter, 2010) as well as the commitment to that decision (e.g. Hupe, Lamirel, & Lorenceau, 2009; Einhauser et al., 2010). However, it is still unclear which specific elements of the decision-making process cause the pupil to dilate.

Recently, de Gee, Knapen, and Donner (2014) sought to link decision-related pupil dilations to the time course of a decision, the decision outcome, as well as participants' own personal response bias towards a decision (i.e. liberal or conservative bias) during a protracted visual discrimination task. In this task, participants had to detect the presence or absence of a visual target within a noisy, static background as fast as possible. The target was a low contrast vertical grating which was superimposed onto the background noise. Half of the trials contained the target and the other half of trials did not contain the target. Responding "present" when the target was present was a hit. Responding "present" when the target was absent was a false alarm. Responding "absent" when the target was present was a miss and responding "absent" when the target was absent was a correct rejection. Thus, this task was conceptually similar to a standard recognition memory task as both tasks require participants to discriminate between signal-present and signal-absent trials, resulting in four decision outcomes (i.e. hits, false alarms, misses, and correct rejections). The key difference, though, is that this is not a memory task, but rather a perceptual task. Researchers have shown that the pupil dilates more for strong memories than for weak memories, such that a large pupil is more likely to indicate a hit than a false alarm (e.g. Otero et al., 2011), but will that same pattern emerge in a memory-less visual discrimination task? In other words, will the pupil dilate more when the target is more visible than when it is less visible and, if so, will hits elicit larger pupils than false alarms? If the pupil does indeed reflect the strength of participants' memory for an item, then

perhaps, in a memory-less discrimination task, the pupil will not dilate more for hits than false alarms.

They found that, relative to a pre-trial baseline, the pupil dilated during the time course of the decision-making process as well as the commitment to the final "present" or "absent" decision, consistent with previous findings (e.g. Einhauser et al., 2010). However, they also found that the pupils dilated more when participants said "present" than when they said "absent," regardless of whether the target was actually present or absent. That is, the pupil did not dilate more for hits than for false alarms, but did dilate more when participants believed that the target was present than when they believed that the target was absent (i.e. hits and false alarms elicited larger pupil dilation than misses and correct rejections). Lastly, the strength of this pupil choice effect seemed to depend on how conservative participants were in their decision to say "present." de Gee et al. split participants into "liberal" and "conservative" subgroups based on the median response criterion of the group. Conservative leaning participants exhibited a very strong pupil choice effect, whereas no effect was observed in the liberal subgroup. de Gee et al. argue that conservative leaning participants were often expecting to say "absent" to most trials, but were surprised when they perceived a strong enough signal to warrant a "present" response. Several recent studies have linked pupil dilation to surprise about perceptual targets (e.g. Hakerem & Sutton, 1966; Privitera, Renninger, Carney, Klein, & Aguilar, 2010) and other behaviorally relevant events (Preuschoff, t Hart, & Einhauser, 2011). Together, these findings show that the pupil not only dilates more when participants say "present" than when they say "absent" but that the strength of this pupil choice effect depends on the participants' liberal or conservative responding.

The Present Experiment

The confidence an eyewitness expresses in their identification decision is a strong predictor of suspect identification accuracy (e.g. Mickes, 2015; Wixted et al., 2016). This means that if an eyewitness makes an identification with high confidence, it is highly likely that the identified suspect is the perpetrator (i.e. a correct ID), whereas if an eyewitness makes an identification with low confidence (e.g. if the eyewitness guesses

that the suspect is the perpetrator), it is increasingly likely that the suspect ID is error prone (i.e. a false ID). Still, there may be other markers of suspect identification accuracy that may further help differentiate a correct ID from a false ID. For example, the way in which eyewitnesses visually scan the perpetrator's face may be different to the way they scan an innocent suspect's face (Mansour et al., 2009; Flowe & Cottrell, 2011), and this information may help differentiate a correct ID from a false ID. In addition, the pupil may dilate more when eyewitnesses make a correct ID than when they make a false ID. The present experiment sought to determine which potential marker of suspect identification accuracy is best. In this experiment, participants studied and attempted to recognize a list of criminal faces. Participants' visual behavior and pupil dilations were measured while making recognition decisions. Participants' confidence in their recognition decisions was also measured in order to conduct CAC analysis (Mickes, 2015).

Methods

Participants

Participants were recruited from Royal Holloway, University of London and completed the experiment in exchange for £10 payment (N = 16; 10 female, 6 male; age in years: M = 20.14; SD = 2.60).

Materials

Participants studied a list of faces gathered from the Florida Department of Corrections Offender Network (http://www.dc.state.fl.us/AppCommon/). This database was used because it helped collect many face images of good quality rather than having to do with the fact that these individuals were criminals. The descriptive details of the criminals included: male, 20-30 years old, white, height of 5'10"-6'2", brown or black hair with no facial hair and no distinguishing features. From here, 60 faces that matched these descriptions were collected and grey-scaled. A mask of each face was made using Adobe Photoshop CS. Small areas (20 X 20 pixels) of each face were randomly mixed so that the image no longer resembled a face, but retained the same level of luminance as the original face image. This mask could serve as a useful baseline when measuring pupil dilations during recognition judgments.

Apparatus and Pupil Measurement

The faces were displayed on a 1024 × 768 pixel CRT monitor (60 Hz refresh rate) at a distance of 68 cm, which was sustained with the use of a table-mounted headrest. Eye movements and maximum pupil size were recorded monocularly (right eye) with an SR Research EyeLink 1000, with a sampling rate of 500 Hz. Due to constant fluctuation in pupil size over time, and variation between individuals, a pupil dilation ratio was calculated. The maximum pupil size from the first 250 ms period of each trial was used as the baseline value and compared with the maximum pupil size from the latter 2750 ms period of each trial. Otero et al. (2011) used this method to calculate the pupil dilation ratio, but others have taken the maximum pupil size value when the item was presented and divided that by the maximum pupil size value when the mask was presented (e.g. Heaver & Hutton, 2011).

Procedure

Following consent, participants were instructed that they were going to take part in a memory test while their eyes were being tracked and their pupils were being recorded. In order to record the changes in pupil size, participants placed their head on top of a headrest which kept their eyes stationary throughout the experiment. In order to respond, participants used a keyboard which was placed slightly below eye level. Before beginning the actual experiment, participants practiced responding on the keyboard while they were positioned on the headrest.

After the practice trials, participants were then calibrated on the eye tracker and began studying the list of faces. During the study phase, 30 faces, randomly selected from the pool of 60 faces, were shown for three seconds each. A mask was briefly shown after each face. After the study phase, participants were recalibrated and began the test phase. During the test phase, old and new faces were randomly intermixed and presented one at a time for an old or new recognition decision. Participants then made a recognition decision using a 1 – 6 confidence scale. Responses 1 – 3 indicated that the face was new with 1 being absolutely certain the face was new and 3 being a guess that the face was new. Responses 4 – 6 indicated that the face was old with 6 being absolutely certain the

face was old and 4 being a guess that the face was old. Confidence was collected in order to conduct CAC analysis. Participants were recalibrated with a one-point calibration after each face was shown. After completing the experiment, participants were debriefed and were paid for their participation. Note that an old/new memory task was used in order to measure changes in the average pupil dilation. An eyewitness identification task would be less suitable because changes in pupil size are very noisy.

Results

Descriptive Analysis

The total number of target-present and target-absent trials as well as the number of hits, false alarms, misses, and correct rejections for every level of confidence are shown in Table 1. Next, we see whether participants' confidence in their recognition decision, their visual behaviour while making a recognition decision, or their pupillary responses to the list of criminal faces can be used to differentiate hits from false alarms.

Table 1

Response frequencies for every recognition decision outcome are displayed for each level of confidence.

		Target-Present		Target-Absent	
		Hits	Misses	False Alarms	Correct Rejections
Said Present	6	113		15	
	5	125		61	
	4	93		64	
Said Absent	3		48		69
	2		71		156
	1		29		116

Note: Responses 4, 5, and 6 indicate that the participant believed the face was present on the study list and responses 1, 2, and 3 indicate that the participant believed the face was not present on the study list.

CAC Analysis

Recognition decisions that received an old response were used to perform CAC analysis. Old responses 4, 5, and 6 refer to low confidence, medium confidence, and high confidence, respectively. The CAC curves in Figure 1 show that as participants recognized a criminal face with increasing confidence, it was increasingly likely that the recognition decision was a hit rather than a false alarm. This means that recognition decisions made with high confidence are highly reliable and recognition decisions made with lower confidence are increasingly error prone. This finding shows, once again, that confidence can be a strong predictor of accuracy – a point which recognition memory theorists (e.g. Egan, 1958; Wixted & Mickes, 2010) and eyewitness identification researchers appreciate (Mickes, 2015; Wixted et al., 2016).

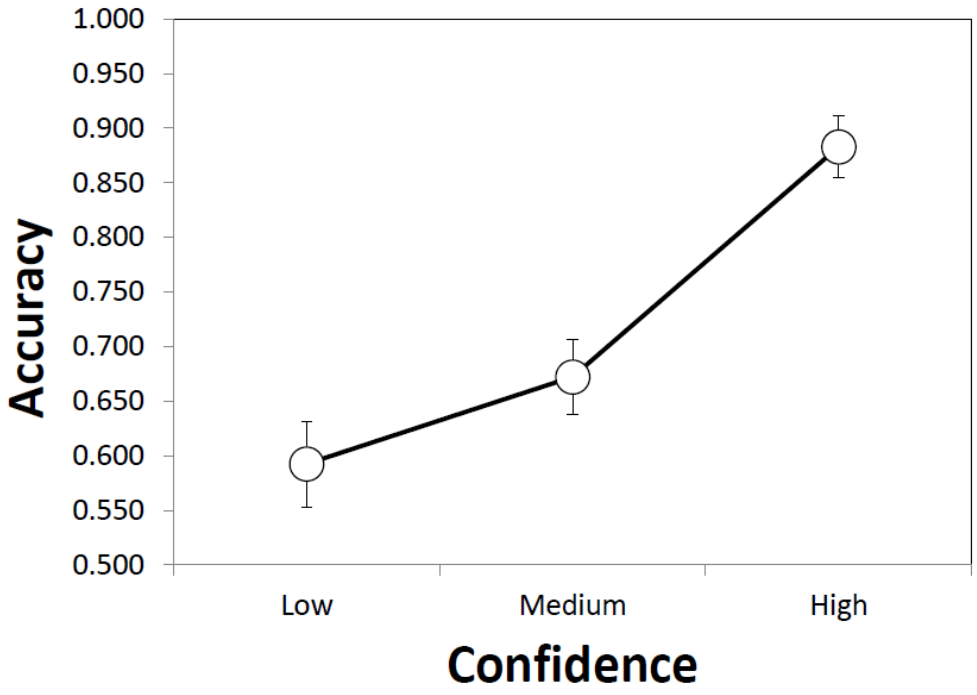

Figure 1. Confidence-accuracy characteristic (CAC) curve for the recognition memory data. The bars represent standard error bars.

Visual Behaviour

Number of Fixations

When participants recognized a face as old, did they fixate more on the correctly recognized faces (i.e. the hits) or the falsely recognized faces (i.e. the false alarms)? Previous studies have found that when participants identify a suspect from a lineup, participants fixate on the falsely identified innocent suspect (i.e. the false IDs) more than the correctly identified perpetrator (i.e. the correct IDs; e.g. Mansour et al., 2009; Flowe & Cottrell, 2011). In this experiment, participants did fixate on the falsely recognized faces (M = 7.4, SD = 2.1) more than the correctly recognized faces (M = 7.2, SD = 2.0), but that difference was not significant, $t(469) = .39$, $p = .70$.

First Fixation Dwell Time

Previous studies have also found that when a suspect is identified from a lineup, participants spend more time first fixating on the correctly identified perpetrator than the falsely identified innocent suspect (e.g. Flowe & Cottrell, 2011). In this experiment, participants spent more time first fixating on the falsely recognized faces (M = 319ms, SD = 300ms) than the correctly recognized faces (M = 300ms, SD = 332ms), but that difference was not significant, $t(469) = .60$, $p = .55$.

Pupil Dilations

Pupil Dilations and Recognition Memory

To determine whether participants' pupils dilated more for old faces (i.e. targets) than for new faces (i.e. lures), the average pupil dilation ratios (PDR) for targets and for lures was compared. Although the average PDR was larger for targets (M = 1.13, SD = .15) than for lures (M = 1.11, SD = .16), that difference was not significant, $t(958) = 1.66$, $p = .09$ (see Figure 2). What is of most interest in this experiment, however, is whether there is a difference in the average PDR for correctly recognized targets (i.e. hits) and for falsely recognized lures (i.e. false alarms). The average PDR for hits (M = 1.14, SD = .16) was greater than the average PDR for false alarms (M = 1.11, SD = .20), but that difference was also not significant, $t(469) = 1.47$, $p = .14$.

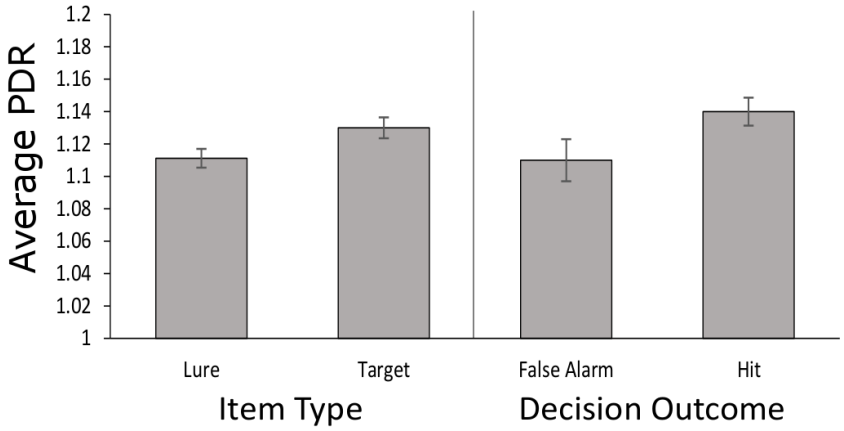

Figure 2. Average pupil dilation ratios collapsed across target and lure items (A) and average pupil dilation ratios for every hits and false alarms (B).

Pupil Dilations and Recognition Decisions

Next, we compared the average PDR for faces that were recognized as old (M = 1.14, SD = .16) with the average PDR for faces that were determined to be new (M = 1.11, SD = .13), and found that difference to be significant, $t(958) = 3.05$, $p = .002$ (see Figure 3A). The PDR was then calculated for each level of confidence. When participants recognized a face as old (i.e. responses 4 – 6), the PDR increased as confidence increased (see Figure 3B). Specifically, the average PDR for faces recognized with high confidence (M = 1.17, SD = .19) was higher than the average PDR for faces recognized with medium confidence (M = 1.14, SD = .14) which, in turn, was higher than the average PDR for faces recognized with low confidence (M = 1.12, SD = .15). The average PDR for faces recognized with high confidence (i.e. 6 responses), medium confidence (i.e. 5 responses), and low confidence (i.e. 4 responses) were compared in a one-way repeated measures ANOVA, which showed a main effect of confidence response, $F(2, 469) = 3.9$, MSE = 0.097, $p = .02$. Subsequent t-tests revealed that participants' pupils dilated significantly more to faces recognized with high confidence (M = 1.17, SD = .19) than to faces recognized with low confidence (M = 1.12, SD = .15, $t(283) = 2.56$, $p = .01$). There was

also a trend for participants' pupils to dilate more to faces recognized with high confidence (M = 1.17, SD = .15) than to faces recognized with medium confidence (M = 1.14, SD = .14; $t(312) = 1.77$, $p = .08$). Thus, participants' pupils dilated more when participants recognized a face as old (i.e. responses 4 – 6) than when they determined a face to be new (i.e. responses 1 – 3), and the more certain they were, the larger their pupils dilated (i.e. 6 responses elicited greater pupil dilations than 5 responses which elicited greater pupil dilations than 4 responses).

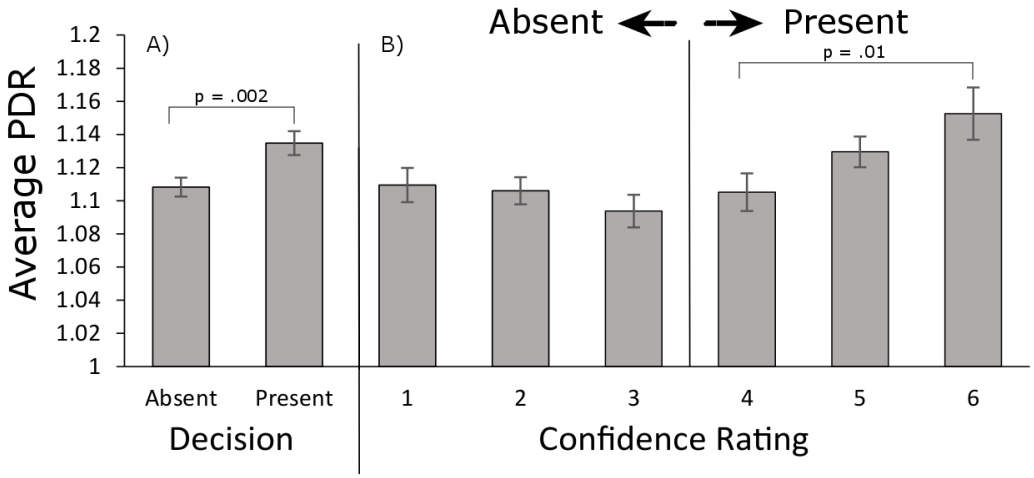

Figure 3. Average pupil dilation ratios collapsed across present and absent decisions (A) and average pupil dilation ratios for every confidence response (B).

General Discussion

Judges and jurors often rely on eyewitness identifications to convict a suspect (e.g. Loftus, 1974). This is problematic when innocent suspects are convicted and later exonerated after their innocence has been established. Eyewitness identification researchers have explored several markers of eyewitness identification accuracy to help differentiate a correct ID from a false ID. These include the confidence the eyewitness reports in their identification decision (Wells & Murray, 1983), the time it takes for the eyewitness to make an identification (Sporer, 1992), and the accuracy of the eyewitnesses'

description of the perpetrator (Pigott & Brigham, 1985). Of these three markers, the confidence the eyewitness reports at the time of their initial identification is arguably the best (Wixted et al., 2016). However, some research suggests that participants scan the face of an innocent suspect differently than how they scan the face of a perpetrator (Flowe & Cottrell, 2011). Research also suggests that participants' pupils will dilate more when making a correct ID than when making a false ID. In this experiment, these three potential markers of eyewitness identification accuracy were explored and compared.

Participants' confidence in their recognition decision was found to be a strong predictor of the accuracy in their decision. If participants were highly confident in their recognition decision, then they were likely to be correct 88% of the time. If participants were less confident, then they were less likely to be correct. Meanwhile, there was no significant difference in the way participants scanned a face when making a hit or a false alarm, and there was no significant difference in participants' pupil dilations when making a hit or a false alarm. Participants' pupils did dilate more when participants believed the face was a target than when they believed the face was a lure, regardless of whether the face was, in fact, a target or a lure. The extent to which their pupils dilated reflected their certainty in that decision. These results are limited because the task used in this experiment was a standard old/new recognition task, rather than an eyewitness identification task. It may be the case that participants' pupils dilate more when making a correct ID than a false ID from a lineup. Barring this limitation, these results suggest 1) there is little use in examining participants' visual behavior and pupil dilations when making a suspect ID and, rather 2) it is extremely useful to measure participants' confidence in their recognition decisions.

Chapter 7

General Discussion

Eyewitness identification researchers continue to have a growing impact on criminal justice systems throughout the world, most notably in the US, by consulting with policymakers or informing the court (i.e. judges and jurors) through expert testimony. Their efforts to shape public policy have mostly been directed at improving aspects of the identification process with the goal to implement procedures that maximize discriminability (e.g. Technical Working Group, 1999). However, eyewitness identification researchers have relied on poor measures of discriminability (i.e. the DR) and, because of this, have encouraged policymakers to adopt procedures that actually reduce discriminability (Gronlund et al., 2014; Gronlund, Wixted, Mickes, & Clark, 2015; Mickes et al., 2012; Wixted & Mickes, 2012). For example, the sequential lineup procedure, which has been repeatedly shown to yield worse discriminability than the traditional simultaneous lineup (e.g. Mickes et al., 2012), has been adopted by approximately 30% of police jurisdictions in the US (Police Executive Forum, 2013). In court, factors that reduce discriminability are believed by many experts to reduce the reliability of an eyewitness identification (Kassin et al., 1989; Kassin et al., 2001).

However, factors that reduce discriminability may not necessarily reduce reliability. In fact, in some cases where discriminability has been reduced, reliability might actually improve (Mickes, 2015). Factors that are believed to reduce discriminability include the verbal overshadowing effect (Schooler & Engstler-Schooler, 1990) and the weapon focus effect (Loftus et al., 1987). I have re-examined these eyewitness phenomena using two analytic techniques recently introduced to the eyewitness identification field that measure discriminability and the reliability of a suspect identification: ROC analysis and CAC analysis, respectively. In total, this thesis consisted of eight empirical studies, each of which highlight the importance of using ROC analysis to measure discriminability and CAC analysis to measure reliability. Failing to measure discriminability and reliability properly could lead to poor public policy and poor trial outcomes, ultimately sending more innocent suspects to prison and more guilty suspects back to the streets.

Chapter 3

Experiment 1 and 2

The United States and the United Kingdom have responded simlarly to reports of erroneous eyewitness identifications (e.g. Devlin, 1976; Technical Working Group, 1999). Both the US and the UK require a lineup to consist of one suspect, who may either be guilty or innocent, and several other fillers. However, the most common lineup procedure in the US is the simultaneous lineup procedure, which consists of six front-face photos presented to the eyewitness simultaneously (Police Executive Forum, 2013). Whereas, in England and Wales, the standard lineup procedure, dictated by the Police and Criminal Evidence Act (PACE), consists of short video clips of nine lineup members presented to the eyewitness in sequence. That entire sequence is shown twice before an identification decision is made (PACE, 1984). Thus, there are several key differences between the US and UK lineup procedures and each of these differences may impact eyewitnesses' discriminability and reliability. A direct comparison between the US lineup and the UK lineup was conducted in order to determine which lineup procedure yields greater discriminability and greater reliability.

ROC Analysis

Because the overall correct and false ID rates were lower for the UK lineup (and could, therefore, mean a shift in response bias, not a difference in discriminability), an analysis of the full ROC was conducted. The US lineup yielded a higher ROC and thus, greater discriminability, than the UK lineup. Because of the myriad of differences between the two lineup procedures, it is difficult to pinpoint exactly why discriminability was greater for the US lineup. I would argue that this finding is yet another example of the often replicated difference between sequential and simultaneous lineups, which often favors the latter (e.g. Carlson & Carlson, 2014; Dobolyi & Dodson, 2013; Gronlund et al., 2012; Mickes et al., 2012).

CAC Analysis

Reliability was assessed by conducting CAC analysis. The US lineup yielded a higher CAC than the UK lineup across low, medium, and high levels of confidence. CAC analysis showed that suspect IDs from the US lineup were more likely to be correct IDs than suspect IDs from the UK lineup, meaning that suspect IDs from the US lineup were more reliable. Again, it is not possible to pinpoint exactly why participants in the UK lineup were less reliable, but a possible explanation for this finding is that participants in the UK lineup failed to appreciate the difficulty in identifying the perpetrator from a sequential lineup presentation and were, thus, worse at determining their accuracy for their suspect IDs. A sequential lineup presentation is purposefully designed to limit the participant's ability to compare lineup members. Wixted and Mickes (2014) argue that this limitation reduces discriminability, but participants may not realize that they are disadvantaged. Because participants are unaware of this disadvantage, their confidence judgments may be less reflective of their suspect ID accuracy. Whatever the reason for this difference, these results underscore the importance of conducting CAC analysis because these results are informative to triers of fact.

Chapter 4

Experiment 3 and 4

An eyewitness to a crime often needs to provide an accurate and detailed description of the perpetrator in order for the police to locate and apprehend the guilty rather than an innocent suspect (Mickes, 2016). In a series of studies, Schooler and Engstler-Schooler (1990) showed that describing the perpetrator can cause a slight, but significant, reduction in the correct ID rate from a lineup. This effect is called "verbal overshadowing" and, although this effect has been replicated many times (Dodson et al., 1997; Fallshore & Schooler, 1995; Ryan & Schooler, 1998; Schooler et al., 1996), many have not found an effect (e.g. Lovett et al., 1992; Memon & Bartlett, 2002; Yu & Geiselman, 1993). The first registered replication report (Alogna et al., 2014), a concerted effort of 31 independent laboratories, attempted to replicate the fourth experiment from Schooler and Engstler-Schooler (1990). Participants in this experiment watched a video

of a mock crime (a bank robbery) and, immediately after the video, either described the perpetrator or engaged in a control task. Participants then took part in a 20 minute distractor task before attempting to identify the perpetrator from a perpetrator-present lineup. A small, but significant verbal overshadowing effect was found. Afterwards, 22 of those laboratories attempted to replicate the first experiment from Schooler and Engstler-Schooler (1990). This time, participants described the perpetrator (or completed a control task for the allotted time) immediately before viewing the lineup. The meta-analytic effect across the 22 studies was much larger (i.e. a much larger reduction in the correct ID rate).

Despite replicating the original report (Alogna et al., 2014), because there was no way to measure false ID rates as the perpetrator was always present in the lineup, the effect verbalization has on identification performance remains unclear (Mickes & Wixted, 2015). That is, it is not clear whether describing the perpetrator impacts discriminability or response bias. To clarify this issue, we conducted a direct replication experiment that included perpetrator-absent lineups, which allows us to measure false ID rates. With both correct and false ID rates, ROC analysis can then be conducted.

If verbal overshadowing reduces discriminability (producing a lower ROC) and not just a conservative shift in responding, Alogna et al. (2014) wrote that suspect IDs admitted as evidence in court "…should be weighted less if the witness had provided a description earlier" (p. 557). Mickes (2015) made the point that results from ROC analyses matter to some decision-makers (e.g., Police Chiefs and policymakers), but another analysis matters more to other decision-makers (e.g., judges and jurors). Because ROC analysis does not measure the reliability of a suspect ID, a more appropriate analysis for judges and jurors is CAC analysis. To test if confidence and accuracy are related at all levels of confidence for both verbal and control conditions, CAC analysis was also conducted.

ROC Analysis

ROC analysis showed that participants who had described the perpetrator were more conservative in making an identification decision than control participants. Lewandowsky (2004) made the point that, because of the noticeable difficulty participants

experience when attempting to describe the perpetrator, participants may believe that their memory for the perpetrator is poor. Participants may therefore require a stronger feeling of recognition in order to make an identification from the lineup. This causes participants to shift their response criteria in a conservative direction. ROC analysis also revealed no significant difference in discriminability between control participants and participants who had described the perpetrator immediately after watching the mock crime video (i.e. 20 minutes before identification). However, there was a significant reduction in discriminability when participants described the perpetrator 20 minutes after watching the mock crime video (i.e. immediately before identification). This reduction in discriminability may have been caused by participants describing the perpetrator using non-diagnostic rather than diagnostic details and then relying on those non-diagnostic details to make an identification (Wixted & Mickes, 2014). A content analysis of the verbal descriptions in both experiments is consistent with this hypothesis. Participants who had described the perpetrator immediately after watching the mock crime video used significantly more diagnostic words in their description than participants who described the perpetrator 20 minutes later. These participants likely relied on those diagnostic details when making an identification, whereas participants who had described the perpetrator 20 minutes after watching the mock crime video likely relied on non-diagnostic details to make an identification.

CAC Analysis

Reliability was compared for participants who had described the perpetrator and control participants by conducting CAC analysis. In both experiments, there was no difference in reliability for high confidence suspect IDs.

Experiment 5

Discriminability should improve when participants rely on diagnostic details to make an identification. This is because diagnostic details are unique to the perpetrator. On the other hand, relying on non-diagnostic details to make an identification should reduce discriminability, as non-diagnostic details are shared by the innocent and guilty suspect (i.e. they are not unique to the perpetrator). This experiment sought to manipulate the

amount of diagnostic details participants use in their description of the perpetrator. Participants either described the specific features of the perpetrator, the general features of the perpetrator, or described a set of unrelated items. Participants who describe the specific features of the perpetrator should yield greater discriminability than control participants, as they are describing features that are unique to the perpetrator. Participants who describe the general features of the perpetrator should yield worse discriminability than control participants, as they are describing features that are likely shared between the perpetrator and the innocent suspect.

ROC Analysis

The general feature condition and the specific feature condition yielded a higher ROC than the control condition, but there was no significant difference in discriminability among the three description groups. Participants in the control condition provided much longer descriptions than participants in the general feature and specific feature conditions. Although instructed to describe the specific and general features of the perpetrator, the relative lack of descriptive details that participants provided indicates that participants had difficulty describing these features. Participants in the control condition therefore had more description-related interference than participants in the other two conditions, which might explain why the general and specific feature condition ROCs are higher, though not significantly higher, than the control condition ROC.

CAC Analysis

As in the previous two experiments, participants who had described the perpetrator were just as reliable as participants who had not previously described the perpetrator. Although researchers have argued that suspect IDs should be given less weight in court if an eyewitness had previously described the perpetrator (e.g. Alogna et al., 2014), CAC results from these three experiments show that participants who had described the perpetrator can still provide reliable high confidence suspect IDs and are just as reliable as participants who had not previously described the perpetrator. In the US, triers of fact are mainly concerned with the reliability for high confidence suspect IDs as low confidence suspect IDs are less likely to be admitted as evidence in court. However, in the

UK, the confidence an eyewitness reports in their identification decision is not collected (PACE, 1984). When an eyewitness makes a low confidence suspect ID, they are essentially telling investigators that they are likely making an error (and they are indeed error prone at low levels of confidence). Neglecting this information may lead to many low confidence suspect IDs being admitted in court.

Chapter 5

Experiment 6 and 7

If a perpetrator is armed during a crime, an eyewitness might focus on the weapon the perpetrator is holding at the expense of focusing on the perpetrator's face (e.g. Loftus et al., 1987). This is called the "weapon focus" effect. An early meta-analysis showed that the presence of a weapon during a crime can cause a slight, but significant, reduction in the correct ID rate (Steblay, 1992). A subsequent meta-analysis (Fawcett et al., 2011) showed that the presence of a weapon causes a small, but significant, reduction in the proportion of correct responses. A reduction in the correct ID rate, or a reduction in proportion correct, do not indicate that the weapon caused a reduction in discriminability as it could be the case that the presence of a weapon caused participants to become more conservative in making an identification. To determine whether the presence of a weapon during a crime causes a reduction in discriminability or a conservative shift in response bias, the correct ID rate *and* the false ID rate must both be measured.

There are ten studies which have measured both the correct and false ID rate. Of these ten studies, four have not found a significant difference in discriminability between the weapon present and weapon absent conditions (Carlson et al., 2016; Cutler & Penrod, 1988; Cutler et al., 1987b; Cutler et al., 1986). Of the six that have found a significant difference in discriminability between weapon present and weapon absent conditions (Carlson & Carlson, 2012; 2014; Carlson et al., 2016; Cutler et al., 1987a; Erickson et al., 2014; O'Rourke et al., 1989), two have used stimuli that display a larger and brighter image of the perpetrator's face in the weapon absent video (Carlson & Carlson, 2012; 2014). Still, a recent survey of eyewitness experts agree with the statement that "The presence of a weapon impairs an eyewitness's ability to accurately identify the

perpetrator's face" and 77% would be willing to testify to that effect in court (Kassin et al., 2001). We conducted two experiments which used a various set of crimes, perpetrators, and weapons in order to determine 1) whether the presence of a weapon during a crime causes a reduction in discriminability and 2) whether participants who had previously seen the perpetrator carrying a weapon are also less reliable.

ROC Analysis

A perceptual analysis found no significant difference in the size or brightness of the perpetrator's face between the weapon absent and weapon present videos. Thus, the data from the eight weapon absent videos and the eight weapon present videos were collapsed into an overall weapon present and weapon absent condition. ROC analysis showed no significant difference in discriminability between the overall weapon present condition and the overall weapon absent condition. However, there appeared to be a trend towards a weapon focus effect in a few of the videos, meaning that discriminability was higher, although not significantly higher, in the weapon absent condition than in the weapon present condition. Perhaps, if more data were collected, this trend towards a weapon focus effect would become significant. Two videos which showed a trend towards a weapon focus effect were used as stimuli in Experiment 2. After recruiting more participants, ROC analysis, again, revealed no significant difference in discriminability between the weapon present condition and the weapon absent condition.

CAC Analysis

Next, we conducted CAC analysis to determine whether the presence of a weapon during a crime caused participants to become less reliable. Across both experiments, participants in the weapon present condition were just as reliable across medium and high levels of confidence as participants in the weapon absent condition. In both groups, as participants' confidence increased so too did their accuracy in their suspect identification. Thus, despite what experts may testify in court about the deleterious effects a weapon may have on suspect identification accuracy, an eyewitness who identifies a suspect with high confidence is highly reliable, regardless of whether the perpetrator held a weapon during the crime.

Chapter 6

Experiment 8

In court, triers of fact tend to place a great deal of faith in the reliability of eyewitness identifications (Loftus, 1974). However, even honest eyewitnesses may accidentally identify an innocent suspect. Eyewitness identification researchers have found several markers of suspect identification accuracy that distinguish, to some degree, correct IDs from false IDs. These include the eyewitnesses' confidence in the identification decision (e.g. Mickes, 2015), the time it takes for an eyewitness to make an identification decision (e.g. Sporer, 1992), and the visual behavior used while scanning the faces of perpetrators and innocent suspects (e.g. Flowe & Cottrell, 2011). However, research from the basic recognition memory literature suggests that participants' pupils may dilate more when they make a correct ID than when they make a false ID (e.g. Otero et al., 2011). This has previously been unexplored in an eyewitness identification task. In this experiment, participants studied and attempted to recognize a list of criminal faces while their pupils were being measured and their eyes were being tracked. Confidence in their recognition judgments were also recorded in order to conduct CAC analysis. Confidence has been shown to be a strong predictor of suspect ID accuracy (e.g. Mickes, 2015). The purpose of this experiment is to compare which potential marker of suspect ID accuracy is best.

CAC Analysis

In order to determine whether confidence could serve as a marker of suspect ID accuracy, CAC analysis was conducted (Mickes, 2015). Participants who were highly confident in their recognition judgments were also highly accurate. Whereas, participants who were less confident in their recognition judgments were less accurate. This means that the confidence an eyewitness reports in their identification decision can serve as a strong predictor of their accuracy in their identification. However, there is still room for improvement. On a few occasions participants made a high confident false alarm. There may be other markers available that can better differentiate a hit from a false alarm.

Visual Behaviour

Previous studies have found that correctly identified targets received fewer fixations than falsely recognized lures (e.g. Flowe & Cottrell, 2011; Mansour et al., 2009). However, there was no significant difference in the number of fixations between hits and false alarms in this experiment. Previous studies have also found that correctly identified targets are first fixated on for a longer duration than falsely recognized lures (e.g. Flowe & Cottrell, 2011). Although hits were first fixated on for a longer duration than false alarms, there was no significant difference between the two decision outcomes.

Pupil Dilations

In the past decade, several recognition memory studies have found a pupil old/new effect, such that participants' pupils dilate more for targets than for lures (e.g. Heaver & Hutton, 2011). Otero et al. (2011) found that when participants say that they recognize an item, participants' pupils dilate more for hits than for false alarms. In a conceptually similar "memory-less" task, participants pupils dilated to the same extent for hits and for false alarms (de Gee et al., 2014). Thus, participants' pupils seem to contain information about the accuracy of their memories as they dilate to a greater extent for hits than for false alarms. In this experiment, there was a trend towards a pupil old/new effect, but the difference in pupil dilation for targets and lures was not significant. Rather, participants' pupils dilated more when they believed the item was previously studied, regardless of whether the item was a target or a lure. This finding however, is not useful as a marker of suspect ID accuracy.

Limitations and Future Research

There are two overarching limitations to the research conducted in this thesis. First, majority of the experiments were conducted online rather than in the confines of a laboratory. The purpose of conducting the experiments online rather than in a laboratory was to obtain thousands of participants in order to conduct ROC and CAC analyses. However, because the experiments were conducted online, it was difficult to control (or even know) participants' environmental factors such as the background noise, lighting

conditions, viewing angle, and the amount of distractions in the environment. Each of these factors likely varied between participants and might have impacted their memory for the perpetrator. For instance, some participants could have participated in the experiment while a movie was playing in the background or while surfing the internet in another browser. Some could have participated in a dark room, making it harder to see the perpetrator's face. Participants were randomly assigned to conditions, which means that these factors were likely spread evenly across conditions; but, because online experiments are less controlled than lab experiments, the data obtained from online experiments may be less valid than data obtained from lab experiments. This is an important limitation to address, but it is also important to note that several researchers have compared data obtained from online experiments and lab experiments and have found online-based data to be just as valid as lab-based data (e.g., Birnbaum, 2001; Buchanan & Smith 1999; Krantz, Ballard, & Scher, 1997; Reips, 2002). In fact, a somewhat recent survey by Musch and Reips (2000) has found that data from 18 online experiments were highly consistent with data from their twin lab experiments. This suggests that the findings in this thesis will likely replicate in a more controlled laboratory setting. Although, the only way to be sure is to conduct these experiments in the laboratory.

A second limitation, that often plagues the extant eyewitness identification literature, is the realism (or lack thereof) of the mock crime videos. In these experiments, participants watched several third-person point of view videos of various mock crimes. The goal was to have participants feel as though they were witnessing an actual crime, but in what respect does viewing videotapes resemble witnessing a crime? According to Cooper et al. (2002), the events typically seen in the laboratory are not comparable to many actual criminal events. For instance, how often is an eyewitness asked to pay close attention when witnessing a crime, which is routine among experimenters conducting eyewitness identification experiments (e.g. Pickel, 1998; 1999)? Eyewitness identification researchers have conducted field studies to address the generally weak ecological validity of eyewitness laboratory experiments (e.g. Amendola & Wixted, 2015; Valentine, Pickering, & Darling, 2003; Wixted et al., 2016). In general, these field studies have been

useful in determining the validity of eyewitness memory effects and the extent to which those effects occur for actual eyewitnesses.

Given these limitations, eyewitness identification researchers should 1) attempt to replicate the findings from these online experiments in the confines of a well-controlled laboratory and, if those results replicate, should 2) conduct these experiments in the field to determine whether these effects extend to the "real world".

Concluding Statement

The research in this thesis sought to determine whether several popular eyewitness memory phenomena impact discriminability or reliability. I re-examined these effects by using two new statistical analyses which measure discriminability and reliability: receiver operating characteristic analysis and confidence-accuracy characteristic analysis, respectively. Altogether, this research has shown that discriminability and reliability are two separate measures of eyewitness identification "accuracy". Understanding that distinction could help improve police procedures and could help inform triers of fact.

References

Aitken, C. G. G. (1995). *Statistics and the evaluation of evidence for forensic scientists*. Chichester, United Kingdom: Wiley.

Alogna, V. K., Attaya, M. K., Aucoin, P., Bahnik, S., Birch, S.… & Birt, A. R. (2014). Registered replication report: Schooler & Engstler-Schooler (1990). *Perspectives on Psychological Science, 9*, 556–579. http://dx.doi.org/10.1177/1745691614545653

Althoff, R. R., & Cohen, N. J. (1999). Eye-movement-based memory effect: a reprocessing effect in face perception. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*(4), 997.

Althoff, R. R., Cohen, N. J., Zelinsky, G., Selco, S., Poldrack, R., & Church, B. (1999). Eye-Movement-Based Memory Effect : A Reprocessing Effect in Face Perception. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *25*(4), 997–1010.

Amendola, K. L., & Wixted, J. T. (2014). Comparing the diagnostic accuracy of suspect identifications made by actual eyewitnesses from simultaneous and sequential lineups in a randomized field trial. *Journal of Experimental Criminology, 11*(2), 263-284. doi:10.1007/s11292-014-9219-2

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

Anderson, M. C., Bjork, R. A., & Bjork, E. L. (1994). Remembering can cause forgetting: Retrieval dynamics in long-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(5), 1063-1087. doi:10.1037//0278-7393.20.5.1063

Bayes T. (1763). An Essay towards solving a Problem in the Doctrine of Chances. *Philosophical Transactions of the Royal Society*, *53*, 370–418. doi:10.1098/rstl.1763.0053

Bazelon, D. L. (1981). Eyewitness news. *Psychology Today*, 101-106.

Beatty, J. (1982). Task-Evoked Pupillary Responses, Processing Load, and the Structure of Processing Resources. *Psychological Bulletin*, *91*(1), 276–292.

Beatty, J., & Lucero-Wagoner, B. (2000). The pupillary system. *Handbook of Psychophysiology, 2*, 142-162.

Behrman, B. W., & Davey, S. L. (2001). Eyewitness identification in actual criminal cases: An archival analysis. *Law and Human Behavior, 25*(5), 475-491. doi:10.1023/a:1012840831846

Birnbaum, M. H. (2001). *Introduction to behavioral research on the Internet*. Pearson College Division.

Brewer, N., & Palmer, M. A. (2010). Eyewitness identification tests. *Legal and Criminological Psychology, 15*(1), 77-96.

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

Broadhurst, P. (1959). The interaction of task difficulty and motivation: The Yerkes-Dodson law revived. *Acta Psychologica, 16*, 321-338. doi:10.1016/0001-6918(59)90105-2

Brown, C., & Lloyd-Jones, T. J. (2002). Verbal overshadowing in a multiple face presentation paradigm: effects of description instruction. *Applied Cognitive Psychology, 16*(8), 873-885. doi:10.1002/acp.919

Brown, C., & Lloyd-Jones, T. J. (2002). Verbal overshadowing of multiple face and car recognition: effects of within- versus across-category verbal descriptions. *Applied Cognitive Psychology*, *17*(2), 183–201. http://doi.org/10.1002/acp.861

Brown, C., & Lloyd-Jones, T. J. (2003). Verbal overshadowing of multiple face and car recognition: effects of within- versus across-category verbal descriptions. *Applied Cognitive Psychology, 17*(2), 183-201. doi:10.1002/acp.861

Brown, C., & Lloyd-Jones, T. J. (2005). Verbal facilitation of face recognition. *Memory & Cognition, 33*(8), 1442-1456. doi:10.3758/bf03193377

Brown, C., & Lloyd-Jones, T. J. (2006). Beneficial effects of verbalization and visual distinctiveness on remembering and knowing faces. *Memory & Cognition*, *34*(2), 277–286.

Brown, C., Lloyd-jones, T. J., Robinson, M., Brown, C., Lloyd-jones, T. J., Robinson, M., & Brown, C. (2008). Eliciting person descriptions from eyewitnesses : A survey of police perceptions of eyewitness performance and reported use of interview techniques. *European Journal of Cognitive Psychology*, *20*(3), 529–560. http://doi.org/10.1080/09541440701728474

Brown, E., Deffenbacher, K., & Sturgill, W. (1977). Memory for faces and the circumstances of encounter. *Journal of Applied Psychology*, *62*(3), 311.

Buchanan, T., & Smith, J. L. (1999). Using the Internet for psychological research: Personality testing on the World Wide Web. *British Journal of Psychology*, *90*(1), 125-144.

Carlson, C. A., & Carlson, M. A. (2012). A distinctiveness-driven reversal of the weapon-focus effect A Distinctiveness-Driven Reversal of the Weapon-Focus Effect. *Applied Psychology in Criminal Justice*, *8*(1), 36–53.

Carlson, C. A., & Carlson, M. A. (2014). An evaluation of lineup presentation, weapon presence, and a distinctive feature using ROC analysis. *Journal of Applied Research in Memory and Cognition, 3*(2), 45-53.

Carlson, C. A., Dias, J. L., Weatherford, D. R., & Carlson, M. A. (2016). An investigation of the weapon focus effect and the confidence–accuracy relationship for eyewitness identification. *Journal of Applied Research in Memory and Cognition*.

Carlson, C. A., Young, D. F., Weatherford, D. R., Carlson, M. A., Bednarz, J. E., & Jones, A. R. (2016). The Influence of Perpetrator Exposure Time and Weapon Presence/Timing on Eyewitness Confidence and Accuracy. *Applied Cognitive Psychology, 30*(6), 898-910. doi:10.1002/acp.3275

Chin, J. M., & Schooler, J. W. (2008). Verbal Overshadowing and Eyewitness Identification. *Encyclopedia of Psychology and Law*. doi:10.4135/9781412959537.n327

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. (2012). Costs and Benefits of Eyewitness Identification Reform: Psychological Science and Public Policy. *Perspectives on Psychological Science, 7*(3), 279-283. doi:10.1177/1745691612444136

Cooper, B. S., Kennedy, M. A., Hervé, H. F., & Yuille, J. C. (2002). Weapon focus in sexual assault memories of prostitutes. *International journal of law and psychiatry*, *25*(2), 181-191.

Costandi, M. (2013, August 14). Evidence-based justice: Corrupted memory. Retrieved January 31, 2017, from http://www.nature.com/news/evidence-based-justice-corrupted-memory-1.13543

Costanzo, M., Krauss, D., & Pezdek, K. (2007). *Expert psychological testimony for the courts*. Mahwah, NJ: Lawrence Erlbaum Associates.

Craik, F. I., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, *11*(6), 671-684.

Craik, F. I., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General*, *104*(3), 268.

Cutler, B. L., & Fisher, R. P. (1990). Live lineups, videotaped lineups, and photoarrays. *Forensic Reports, 3*, 439–448.

Cutler, B. L., & Penrod, S. D. (1988). Improving the reliability of eyewitness identification: Lineup construction and presentation. *Journal of Applied Psychology, 73*(2), 281.

Cutler, B. L., Berman, G. L., Penrod, S., & Fisher, R. P. (1994). Conceptual, practical and empirical issues associated with eyewitness identification test media. In D. F. Ross, J. D. Read, & M. P. Toglia (Eds.), *Adult eyewitness testimony: Current trends and developments* (pp. 163–181). Cambridge: Cambridge University Press.

Cutler, B. L., Penrod, S. D., & Martens, T. K. (1987). Improving the reliability of eyewitness identification: Putting context into context. *Journal of Applied Psychology, 72*(4), 629-637. doi:10.1037//0021-9010.72.4.629

Cutler, B. L., Penrod, S. D., O'Rourke, T. E., & Martens, T. K. (1986). Unconfounding the effects of contextual cues on eyewitness identification accuracy. *Social Behaviour, 1*(2), 113-134.

Cutler, B. L., Penrod, S. D., Rourke, O., Coward, J., Fleicher, R., Guth, M. K., … Wagman, Z. (1988). Improving the Reliability of Eyewitness Identification : Lineup Construction and Presentation. *Journal of Applied Psychology*, *73*(2), 281–290.

Darley, C. F., & Glass, A. L. (1975). Effects of rehearsal and serial list position on recall. *Journal of Experimental Psychology: Human Learning & Memory, 1*(4), 453-458. doi:10.1037//0278-7393.1.4.453

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

Davis, D., & Follette, W. C. (2002). Rethinking the probative value of evidence: Base rates, intuitive profiling, and the "postdiction" of behavior. *Law and Human Behavior, 26*(2), 133-158. doi:10.1023/a:1014693024962

de Gee, J. W., Knapen, T., & Donner, T. H. (2014). Decision-related pupil dilation reflects upcoming choice and individual bias. *Proceedings of the National Academy of Sciences*, *111*(5), E618-E625.

Deffenbacher, K. A. (1983). The influence of arousal on reliability of testimony. In S. M. A. Lloyd-Bostock & B. R. Clifford (Eds.). *Evaluating witness evidence*. Chichester: Wiley. (pp. 235-251).

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.

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

Dodson, C. S., Johnson, M. K., & Schooler, J. W. (1997). The verbal overshadowing effect: why descriptions impair face recognition. *Memory & Cognition*, *25*(2), 129–139.

Dodson, C. S., Johnson, M. K., & Schooler, J. W. (1997). The verbal overshadowing effect: Why descriptions impair face recognition. *Memory & Cognition, 25*(2), 129-139. doi:10.3758/bf03201107

Duffy, E. (1962). *Activation and behavior*. New York: Wiley.

Dunn, J. C. (2004). Remember-know: a matter of confidence. *Psychological review*, *111*(2), 524.

Easterbrook, J. A. (1959). The effect of emotion on cue utilization and the organization of behavior. *Psychological Review, 66*(3), 183-201. doi:10.1037/h0047707

Egan, J. P. (1958). Recognition memory and the operating characteristic (Tech. Note AFCRC-TN-58–51). Bloomington: Indiana University, Hearing and Communication Laboratory.

Egeth, H. E. (1993). What Do We Not Know About Eyewitness Identification ? *American Psychologist*, *48*(5), 577–580.

Einhauser, W., Koch, C., & Carter, O. (2010). Pupil dilation betrays the timing of decisions. *Frontiers in Human Neuroscience*, *4*, 18.

Einhäuser, W., Stout, J., Koch, C., & Carter, O. (2008). Pupil dilation reflects perceptual selection and predicts subsequent stability in perceptual rivalry. *Proceedings of the National Academy of Sciences*, *105*(5), 1704-1709.

Ellison, K. W., & Buckhout, R. (1981). Psychology and criminal justice (pp. 108-110). New York: Harper & Row.

Ellman, I. M., & Kaye, D. H. (1979). Probabilities and proof: Can HLA and blood test evidence prove paternity? *New York University Law Review, 55*, 1131–1162.

Erickson, W. B., Lampinen, J. M., & Leding, J. K. (2014). The Weapon Focus Effect in Target-Present and Target-Absent Line-Ups: The Roles of Threat, Novelty, and Timing. *Applied Cognitive Psychology, 28*(3), 349-359. doi:10.1002/acp.3005

Fallshore, M., & Schooler, J. W. (1995). Verbal vulnerability of perceptual expertise. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*(6), 1608-1623. doi:10.1037//0278-7393.21.6.1608

Fawcett, J. M., Russell, E. J., Peace, K. A., & Christie, J. (2013). Of guns and geese: a meta-analytic review of the 'weapon focus' literature. *Psychology, Crime & Law, 19*(1), 35-66. doi:10.1080/1068316x.2011.599325

Fenton, N.E., Neil, M. (2011). Avoiding legal fallacies in practice using Bayesian networks. *Australian Journal of Legal Philosophy, 36*, 114–50.

Fienberg, S. E., & Kadane, J. B. (1983). The Presentation of Bayesian Statistical Analyses in Legal Proceedings. *The Statistician, 32*(1/2), 88. doi:10.2307/2987595

Fienberg, S.E. & Finkelstein, M.O. (1996). Bayesian statistics and the law. In *Bayesian Statistics 5*, ed. JM Bernardo, JO Berge, AP Dawid, AFM Smith, pp. 129–46. Oxford, UK: Oxford Univ. Press

Fienberg, S.E. (2011). Bayesian models and methods in public policy and government settings. *Statistical Science, 26*, 212–26.

Finger, K. (2002). Mazes and music: using perceptual processing to release verbal overshadowing. *Applied Cognitive Psychology, 16*(8), 887-896. doi:10.1002/acp.922

Finger, K., & Pezdek, K. (1999). The effect of cognitive interview on face identification accuracy: Release from verbal overshadowing. *Journal of Applied Psychology, 84*(3), 340-348. doi:10.1037//0021-9010.84.3.340

Finkelstein, M. O. (1978). *Quantitative Methods in Law: Studies in the Application of Mathematical Probability and Statistics to Legal Problems*. New York, NY: The Free Press.

Finkelstein, M. O., & Fairley, W. B. (1970). A Bayesian Approach to Identification Evidence. *Harvard Law Review, 83*(3), 489. doi:10.2307/1339656

Fisher, R. P., & Geiselman, R. E. (1992). *Memory-enhancing techniques for investigative interviewing: the cognitive interview*. Springfield, IL, U.S.A.: Thomas.

Flowe, H. (2011). An Exploration of Visual Behaviour in Eyewitness Identification Tests. *Applied Cognitive Psychology*, *25*, 244–254.

Flowe, H. (2011). An exploration of visual behaviour in eyewitness identification tests. *Applied Cognitive Psychology, 25*(2), 244-254.

Flowe, H., & Cottrell, G. W. (2011). An examination of simultaneous lineup identification decision processes using eye tracking. *Applied Cognitive Psychology, 25*(3), 443-451. doi:10.1002/acp.1711

Frazzini, S. F. (1981). Review of eyewitness testimony. *The Yale Review*, *70*, 18-20.

Friedman, R. D. (1986). A close look at probative value. *Boston University Law Review, 66*, 733–759.

Gardner, R. M., Mo, S. S., & Borrego, R. (1974). Inhibition of pupillary orienting reflex by novelty in conjunction with recognition memory. *Bulletin of the Psychonomic Society*, *3*(3), 237-238.

Gardner, R. M., Mo, S. S., & Krinsky, R. (1974). Inhibition of pupillary orienting reflex by heteromodal novelty. *Bulletin of the Psychonomic Society*, *4*(5), 510-512.

Gary, L., & Lindsay, C. L. (1980). On Estimating the Diagnosticity of Eyewitness Nonidentifications. *Psychological Bulletin*, *88*(3), 776–784.

Geiselman, R. E., Fisher, R. P., Firstenberg, I. Hutton, L. A., Sullivan, S. J., Avetissain, I. V., Prosk, A. L. (1984). Enhancement of eyewitness memory: An empirical evaluation of the cognitive interview. *Journal of Police Science and Administration, 12*(1), 74-80.

Geiselman, R. E., Fisher, R. P., Mackinnon, D. P., & Holland, H. L. (1985). Eyewitness Memory Enhancement in the Police Interview : Cognitive Retrieval Mnemonics Versus Hypnosis. *Journal of Applied Psychology*, *70*(2), 401–412.

Glenberg, A., & Adams, F. (1978). Type I rehearsal and recognition. *Journal of Verbal Learning and Verbal Behavior, 17*(4), 455-463. doi:10.1016/s0022-5371(78)90274-8

Glenberg, A., Smith, S. M., & Green, C. (1977). Type I rehearsal: Maintenance and more. *Journal of Verbal Learning and Verbal Behavior, 16*(3), 339-352. doi:10.1016/s0022-5371(77)80055-8

Goldinger, S. D., & Papesh, M. H. (2012). Pupil dilation reflects the creation and retrieval of memories. *Current Directions in Psychological Science*, *21*(2), 90-95.

Goldinger, S. D., He, Y., & Papesh, M. H. (2009). Deficits in cross-race face learning: insights from eye movements and pupillometry. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(5), 1105.

Goldstein, A. G., Chance, J. E., & Schneller, G. R. (1989). Frequency of eyewitness identification in criminal cases: A survey of prosecutors. *Bulletin of the Psychonomic Society, 27*(1), 71-74.

Goodman, S. (1999). Towards Evidence-Based Medical Statistics, 2: The Bayes Factor. *Annals of Internal Medicine*, *130*, 1005-1013.

Goodman, S. (2005). Judgment for Judges: What Traditional Statistics Don't Tell You About Causal Claims. *Brooklyn Law School Review, 15*.

Gourevitch, V., & Galanter, E. (1967). A significance test for one parameter isosensitivity functions. *Psychometrika, 32*(1), 25-33. doi:10.1007/bf02289402

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

Gronlund, S. D., Carlson, C. A., Dailey, S. B., & Goodsell, C. A. (2009). Robustness of the Sequential Lineup Advantage. *Journal of Experimental Psychology. Applied*, *15*(2), 140–152. http://doi.org/10.1037/a0015082

Gronlund, S. D., Mickes, L., Wixted, J. T., & Clark, S. E. (2015). Conducting an Eyewitness Lineup: How the Research Got It Wrong. *Psychology of Learning and Motivation,* 1-43. doi:10.1016/bs.plm.2015.03.003

Gronlund, S. D., Wixted, J. T., & Mickes, L. (2014). Current Directions in Psychological Science. *Current Directions in Psychological Science*, *23*(1), 3–10. http://doi.org/10.1177/0963721413498891

Gronlund, S.D., Carlson, C.A., Neuschatz, J.S., Goodsell, C.A., Wetmore, S.A., Wooten, A., & Graham, M. (2012). *Journal of Applied Research in Memory and Cognition*, *1*, 221-228. doi:10.1037/e571212013-331

Hakerem, G. A. D., & Sutton, S. (1966). Pupillary response at visual threshold. *Nature*.

Hansen, M. (2001). Forensic science: Scoping out eyewitness IDs. *American Bar Association Journal*, *87*, 39.

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*, 1209–1221. http://doi.org/10.1002/acp

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

Heaver, B., & Hutton, S. B. (2010). Keeping an eye on the truth: Pupil size, recognition memory and malingering. *International Journal of Psychophysiology, 77*(3), 306-306.

Hebb, D. O. (1955). Drives and the C. N. S. (conceptual nervous system). *Psychological Review, 62*(4), 243-254. doi:10.1037/h0041823

Henderson, J. M., Williams, C. C., & Falk, R. J. (2005). Eye movements are functional during face learning. *Memory & Cognition, 33*(1), 98-106.

Hess, E. H., & Polt, J. M. (1964). Pupil size in relation to mental activity during simple problem-solving. *Science*, *143*(3611), 1190-1192.

Home Office (2004). Police and Criminal Evidence Act 1984 (s.60 (1)(a), s.60A(1) and s.66(1)). Codes of Practice A-F. Revised Edition. London: The Stationery Office.

Horry, R., Palmer, M. a, & Brewer, N. (2012). Backloading in the sequential lineup prevents within-lineup criterion shifts that undermine eyewitness identification performance. *Journal of Experimental Psychology. Applied*, *18*(4), 346–60. http://doi.org/10.1037/a0029779

Horry, R., Palmer, M. A., & Brewer, N. (2012). Backloading in the sequential lineup prevents within-lineup criterion shifts that undermine eyewitness identification performance. *Journal of Experimental Psychology: Applied, 18*(4), 346-360. doi:10.1037/a0029779

Hulse, L. M., & Memon, A. (2006). Fatal impact? The effects of emotional arousal and weapon presence on police officers' memories for a simulated crime. *Legal and Criminological Psychology, 11*(2), 313-325. doi:10.1348/135532505x58062

Humphreys, M. S., & Revelle, W. (1984). Personality, motivation, and performance: A theory of the relationship between individual differences and information processing. *Psychological Review, 91*(2), 153-184. doi:10.1037//0033-295x.91.2.153

Humphries, J. E., Holliday, R. E., & Flowe, H. D. (2011). Faces in Motion: Age-Related Changes in Eyewitness Identification Performance in Simultaneous, Sequential, and Elimination Video Lineups. *Applied Cognitive Psychology, 26*(1), 149-158. doi:10.1002/acp.1808

Hupé, J. M., Lamirel, C., & Lorenceau, J. (2009). Pupil dynamics during bistable motion perception. *Journal of vision*, *9*(7), 10-10.

Innocence Project (2017). Help us put an end to wrongful convictions! Retrieved February 01, 2017, from http://www.innocenceproject.org/.

Itoh, Y. (2005). The facilitating effect of verbalization on the recognition memory of incidentally learned faces. *Applied Cognitive Psychology*, *19*(4), 421–433. http://doi.org/10.1002/acp.1069

Jackson v. Fogg, 589 F.2d 108 (2d Cir. 1978).

Janisse, M. P. (1977). *Pupillometry: The psychology of the pupillary response*. Halsted Press.

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

Kafkas, A., & Montaldi, D. (2011). Recognition memory strength is predicted by pupillary responses at encoding while fixation patterns distinguish recollection from familiarity. *The Quarterly Journal of Experimental Psychology*, *64*(10), 1971-1989.

Kahneman, D., & Beatty, J. (1966). Pupil diameter and load on memory. *Science*, *154*(3756), 1583-1585.

Kassin, S. M., Ellsworth, P. C., & Smith, V. L. (1989). The "general acceptance" of psychological research on eyewitness testimony: A survey of the experts. *American Psychologist, 44*(8), 1089-1098. doi:10.1037//0003-066x.44.8.1089

Kassin, S. M., Tubb, V. A., Hosch, H. M., & Memon, A. (2001). On the "general acceptance" of eyewitness testimony research: A new survey of the experts. *American Psychologist, 56*(5), 405-416. doi:10.1037/0003-066x.56.5.405

Kaye, D. H., & Koehler, J. J. (2003). The Misquantification of Probative Value. *Law and Human Behavior, 27*(6), 645-659. doi:10.1023/b:lahu.0000004892.94380.88

Kendall, E. (1981). *The phantom prince: my life with Ted Bundy*. Seattle: Madrona.

Kerstholt, J. H., Koster, E. R., & 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

Kitagami, S., Sato, W., & Yoshikawa, S. (2002). The influence of test-set similarity in verbal overshadowing. *Applied Cognitive Psychology, 16*(8), 963-972. doi:10.1002/acp.917

Klobuchar, A., Steblay, N. K. M., & Caligiuri, H. L. (2006). Improving eyewitness identifications: Hennepin County's blind sequential lineup pilot project. *Cardozo Public Law, Policy and Ethics Journal, 2*, 381–414.

Konečni, V. J., & Ebbesen, E. B. (1986). Courtroom testimony by psychologists on eyewitness identification issues: Critical notes and reflections. *Law and Human Behavior, 10*(1-2), 117-126. doi:10.1007/bf01044563

Kramer, T. H., Buckhout, R., & Eugenio, P. (1990). Weapon focus, arousal, and eyewitness memory: Attention must be paid. *Law and Human Behavior, 14*(2), 167-184. doi:10.1007/bf01062971

Krantz, J. H., Ballard, J., & Scher, J. (1997). Comparing the results of laboratory and World-Wide Web samples on the determinants of female attractiveness. *Behavior Research Methods, Instruments, & Computers*, *29*(2), 264-269.

Leippe, M. R. (1995). The case for expert testimony about eyewitness memory. *Psychology, Public Policy, and Law*, *1*(4), 909–959. http://doi.org/10.1037//1076-8971.1.4.909

Levine, F., & Tapp, J. (1973). The psychology of criminal identification: The gap from Wade to Kirby. *University of Pennsylvania Law Review, 5*, 1079-1131.

Lindsay, R. C. (1986). Confidence and accuracy of eyewitness identification from lineups. *Law and Human Behavior, 10*(3), 229-239.

Lindsay, R. C., & Wells, G. L. (1980). What price justice? Exploring the relationship of lineup fairness to identification accuracy. *Law and Human Behavior, 4*(4), 303-313. doi:10.1007/bf01040622

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

Lindsay, R. C., Wells, G. L., & Rumpel, C. M. (1981). Can people detect eyewitness-identification accuracy within and across situations?. *Journal of Applied Psychology*, *66*(1), 79.

Lloyd-Jones, T. J., & Brown, C. (2008). Verbal overshadowing of multiple face recognition: Effects on remembering and knowing over time. *European Journal of Cognitive Psychology*, *20*(3), 456–477. http://doi.org/10.1080/09541440701728425

Loftus, E. F. (1975). Reconstructing memory: The incredible eyewitness. *Jurimetrics Journal, 15*(3), 188-193.

Loftus, E. F. (1979). *Eyewitness testimony*. Cambridge, MA: Harvard University Press.

Loftus, E. F. (1983). Silence is not golden. *American Psychologist, 38*(5), 564-572. doi:10.1037//0003-066x.38.5.564

Loftus, E. F., & Ketcham, K. (1992). *Witness for the defense: the accused, the eyewitness, and the expert who puts memory on trial*. New York: St. Martin's Press.

Loftus, E. F., & Palmer, J. C. (1974). Reconstruction of automobile destruction: An example of the interaction between language and memory. *Journal of Verbal Learning and Verbal Behavior, 13*(5), 585-589. doi:10.1016/s0022-5371(74)80011-3

Loftus, E. F., Loftus, G. R., & Messo, J. (1987). Some facts about "weapon focus.". *Law and Human Behavior, 11*(1), 55-62. doi:10.1007/bf01044839

Loftus, G. R., & Mackworth, N. H. (1978). Cognitive determinants of fixation location during picture viewing. *Journal of Experimental Psychology: Human Perception and Performance, 4*(4), 565-572. doi:10.1037//0096-1523.4.4.565

Lovett, S. B., Small, M. Y., & Engstrom, S. A. (1992, November). The verbal overshadowing effect: Now you see it, now you don't. Paper presented at the annual meeting of the Psychonomic Society, St. Louis, MO.

Luus, C. A., & Wells, G. L. (1994). The malleability of eyewitness confidence: Co-witness and perseverance effects. *Journal of Applied Psychology*, *79*(5), 714.

Maass, A., & Köhnken, G. (1989). Eyewitness identification: Simulating the "weapon effect.". *Law and Human Behavior, 13*(4), 397-408. doi:10.1007/bf01056411

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: a user's guide*. Mahwah, NJ: Lawrence Erlbaum Associates.

Macrae, C. N., & Lewis, H. L. (2002). Do I Know You? Processing Orientation and Face Recognition. *Psychological Science, 13*(2), 194-196. doi:10.1111/1467-9280.00436

Malpass, R. S. (1981). Effective size and defendant bias in eyewitness identification lineups. *Law and Human Behavior, 5*(4), 299-309. doi:10.1007/bf01044945

Mansour, J. K., Lindsay, R. C. L., Brewer, N., & Munhall, K. G. (2009). Characterizing visual behaviour in a lineup task. *Applied Cognitive Psychology, 23*(7), 1012-1026.

Marshall, C. R., & Wise, J. A. (1975). Juror Decisions and the Determination of Guilt in Capital Punishment Cases: A Bayesian Perspective. *Utility, Probability, and Human Decision Making,* 257-269. doi:10.1007/978-94-010-1834-0_15

Maw, N. N., & Pomplun, M. (2004, January). Studying human face recognition with the gaze-contingent window technique. In *Proceedings of the Cognitive Science Society* (Vol. 26, No. 26).

McGeoch, J.A. (1932). Forgetting and the law of disuse. *Psychological Review, 39*, 352–70.

Mecklenburg, S.H. (2006). Report to the legislature of the State of Illinois: The Illinois Pilot Program on double-blind, sequential lineup procedures. Springfield, IL: Illinois State Police.

Meissner, C. A., & Brigham, J. C. (2001). A meta-analysis of the verbal overshadowing effect in face identification. *Applied Cognitive Psychology, 15*(6), 603-616. doi:10.1002/acp.728

Meissner, C. A., Brigham, J. C., & Kelley, C. M. (2001). The influence of retrieval processes in verbal overshadowing. *Memory & Cognition, 29*(1), 176-186. doi:10.3758/bf03195751

Meissner, C. A., Sporer, S. L., & Susa, K. J. (2008). A theoretical review and meta-analysis of the description-identification relationship in memory for faces. *European Journal of Cognitive Psychology, 20*(3), 414-455. doi:10.1080/09541440701728581

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

Melcher, J. M., & Schooler, J. W. (1996). The Misremembrance of Wines Past: Verbal and Perceptual Expertise Differentially Mediate Verbal Overshadowing of Taste Memory. *Journal of Memory and Language, 35*(2), 231-245. doi:10.1006/jmla.1996.0013

Memon, A., & Bartlett, J. (2002). The effects of verbalization on face recognition in young and older adults. *Applied Cognitive Psychology, 16*(6), 635-650. doi:10.1002/acp.820

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

Memon, A., Mastroberardino, S., & Fraser, J. (2008). Münsterberg's legacy: What does eyewitness research tell us about the reliability of eyewitness testimony? *Applied Cognitive Psychology, 22*(6), 841-851. doi:10.1002/acp.1487

Mickes, L. (2015). Receiver operating characteristic analysis and confidence – accuracy characteristic analysis in investigations of system variables and estimator variables that affect eyewitness memory. *Journal of Applied Research in Memory and Cognition*, *4*(2), 93–102. http://doi.org/10.1016/j.jarmac.2015.01.003

Mickes, L. (2016). The Effects of Verbal Descriptions on Eyewitness Memory : Implications for the Real- World The Effects of Verbal Descriptions on Eyewitness Memory : *Journal of Applied Research in Memory and Cognition*, *5*, 270–276. http://doi.org/10.1016/j.jarmac.2016.07.003

Mickes, L., & Wixted, J. T. (2015). On the applied implications of the "verbal overshadowing effect". *Perspectives on Psychological Science*, *10*(3), 400-403.

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. http://doi.org/10.1037/a0030609

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

Mitchell, K. J., Livosky, M., & Mather, M. (1998). The weapon focus effect revisited: The role of novelty. *Legal and Criminological Psychology, 3*(2), 287-303. doi:10.1111/j.2044-8333.1998.tb00367.x

Monahan, J., & Walker, L. (1988). Social science research in law: A new paradigm. *American Psychologist, 43*(6), 465-472. doi:10.1037//0003-066x.43.6.465

Moore, C. (1907). Yellow psychology. *Law Notes, 11,* 125–127.

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior, 16*(5), 519-533. doi:10.1016/s0022-5371(77)80016-9

Muller GE, Pilzecker A. (1900). Experimentelle Beitrage zur Lehre vom Gedachtnis [Experimental contributions to the science of memory]. *Z. Psychol. Erganz. 1*, 1–300.

Münsterberg, H. (1908). *On the witness stand; essays on psychology and crime*. New York: McClure Co.

Musch, J., & Reips, U. D. (2000). A brief history of Web experimenting. (pp. 61-87). San Diego, CA, US: Academic Press, 20, 317 pp. http://dx.doi.org/10.1016/B978-012099980-4/50004-6

National Research Council. (2015). *Identifying the culprit: Assessing eyewitness identification*. National Academies Press.

Navon, D. (1977). Forest before trees: The precedence of global features in visual perception. *Cognitive Psychology, 9*(3), 353-383. doi:10.1016/0010-0285(77)90012-3

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

O'Rourke, T. E., Penrod, S. D., Cutler, B. L., & Stuve, T. E. (1989). The external validity of eyewitness identification research: Generalizing across subject populations. *Law and Human Behavior, 13*(4), 385.

Otero, S. C., Weekes, B. S., & Hutton, S. B. (2011). Pupil size changes during recognition memory. *Psychophysiology*, *48*(10), 1346-1353.

Otero, S., Weeks, B., & Hutton, S. (2006). A novel association between pupil size and recollective experience during recognition memory. In Abstract presented at The Second Biennial Conference on Cognitive Science, St Petersburg, Russia.

Palmer, M. A., Brewer, N., Weber, N., & Nagesh, A. (2013). The Confidence-Accuracy Relationship for Eyewitness Identification Decisions : Effects of Exposure Duration, Retention Interval, and Divided Attention. *Journal of Experimental Psychology. Applied*, *19*(1), 55–71. http://doi.org/10.1037/a0031602

Penrod, S. D., & Cutler, B. L. (1989). Eyewitness Expert Testimony and Jury Decisionmaking. *Law and Contemporary Problems, 52*(4), 43. doi:10.2307/1191907

Penrod, S. D., Fulero, S. M., Itc, B., Cutler, L., Penrod, S. D., Fulero, S. M., & Cutler, B. L. (1995). Expert Psychological Testimony in the United States : A New Playing Field ?*. *European Journal of Psychological Assessment*, *11*(1), 65–72.

People v. McDonald, 690 P.2d 709 (Cal. 1984).

Pickel, K. L. (1998). Unusualness and Threat as Possible Causes of "Weapon Focus". *Memory, 6*(3), 277-295. doi:10.1080/741942361

Pickel, K. L. (1999). The Influence of Context on the “Weapon Focus” Effect. *Law and Human Behavior*, *23*(3), 299–311.

Pigott, M., & Brigham, J. C. (1985). Relationship between accuracy of prior description and facial recognition. *Journal of Applied Psychology, 70*(3), 547.

Pike, G., Brace, N., & Kynan, S. (2002). The visual identification of suspects: Procedures and practice. *Home Office, Briefing Note*, *2*(02).

Police Executive Research Forum (PERF), & United States of America. (2013). National Survey of Eyewitness Identification Procedures in Law Enforcement Agencies.

Postman L, Alper T. 1946. Retroactive inhibition as a function of the time interpolation of the inhibitor between learning and recall. *American Journal of Psychology, 59*, 439–49.

Preuschoff, K., t Hart, B. M., & Einhauser, W. (2011). Pupil dilation signals surprise: Evidence for noradrenaline’s role in decision making. *Frontiers in neuroscience*, *5*, 115.

Privitera, C. M., Renninger, L. W., Carney, T., Klein, S., & Aguilar, M. (2010). Pupil dilation during visual target detection. *Journal of Vision*, *10*(10), 3-3.

Puch-Solis, R., P. Roberts, S. Pope & C. Aitken (2012). PRACTITIONER GUIDE NO 2: Assessing the Probative Value of DNA Evidence, Guidance for Judges, Lawyers,

Forensic Scientists and Expert Witnesses, Royal Statistical Society http://www.rss.org.uk/uploadedfiles/userfiles/files/Practitioner-Guide-2-WEB.pdf

Reips, U. D. (2002). Standards for Internet-based experimenting. *Experimental psychology*, *49*(4), 243.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J., & Müller, M. (2011). PROC: an open-source package for R and S to analyze and compare ROC curves. *BMC Bioinformatics, 12*(1), 77. doi:10.1186/1471-2105-12-77

Roediger, H. L. (1990). Implicit memory: Retention without remembering. *American Psychologist, 45*(9), 1043-1056. doi:10.1037//0003-066x.45.9.1043

Roediger, H.L., Weldon, M. S., & Challis, B. H. (1989). Explaining dissociations between implicit and explicit measures of retention: A processing account. In Tulving, E., Roediger, H. L., & Craik, F. I. *Varieties of memory and consciousness: essays in honour of Endel Tulving*. Hillsdale, NJ: L. Erlbaum Associates.

Rotello, C. M., & Zeng, M. (2008). Analysis of RT distributions in the remember—know paradigm. *Psychonomic Bulletin & Review*, *15*(4), 825-832.

Rotello, C. M., Heit, E., & Dubé, C. (2014). When more data steer us wrong: replications with the wrong dependent measure perpetuate erroneous conclusions. *Psychonomic Bulletin & Review, 22*(4), 944-954. doi:10.3758/s13423-014-0759-2

Rundus, D. (1971). Analysis of rehearsal processes in free recall. *Journal of Experimental Psychology, 89*(1), 63-77. doi:10.1037/h0031185

Ryan, J. D., Hannula, D. E., & Cohen, N. J. (2007). The obligatory effects of memory on eye movements. *Memory, 15*(5), 508-525.

Ryan, R. S., & Schooler, J. W. (1998). Whom do words hurt? Individual differences in susceptibility to verbal overshadowing. *Applied Cognitive Psychology, 12*(7). doi:10.1002/(sici)1099-0720(199812)12:73.3.co;2-m

Satake, E., & Amato, P. P. (1999). Probability and Law: A Bayesian Approach. The Bulletin of International Statistics Instistute 52nd Session, Helsinki, Finland.

Sauer, J., Brewer, N., Zweck, T., & Weber, N. (2010). The effect of retention interval on the confidence–accuracy relationship for eyewitness identification. *Law and Human Behavior, 34*(4), 337-347. doi:10.1007/s10979-009-9192-x

Schooler, J. W. (2002). Verbalization produces a transfer inappropriate processing shift. *Applied Cognitive Psychology, 16*, 989-997.

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

Schooler, J. W., Fiore, S. M., & Brandimonte, M. A. (1997). At a Loss From Words: Verbal Overshadowing of Perceptual Memories. *Psychology of Learning and Motivation,* 291-340. doi:10.1016/s0079-7421(08)60505-8

Schooler, J. W., Ryan, R. S., & Reder, L. M. (1996). The costs and benefits of verbally rehearsing memory for faces. In D. Herrmann, M. K. Johnson, C. McEvoy, C. Hertzog, & P.Hertel (Eds.), *Basic and applied memory: New findings* (pp. 51-65). Hillsdale, NJ: Erlbaum.

Seale-Carlisle, T. M., & Mickes, L. (2016). US line-ups outperform UK line-ups. *Royal Society Open Science*, *3*: 160300. http://dx.doi.org/10.1098/rsos.160300

Shaw, J. I., & Skolnick, P. (1994). Sex differences, weapon focus, and eyewitness reliability. *The Journal of social psychology, 134*(4), 413-420.

Simpson, H. M., & Hale, S. M. (1969). Pupillary changes during a decision-making task. *Perceptual and Motor Skills*, *29*(2), 495-498.

Skaggs EB. (1925). Further studies in retroactive inhibition. *Psychology Monograph*, *34*, 1–60.

Smith, S. M., Lindsay, R. C. L., & Pryke, S. (2000). Postdictors of eyewitness errors: Can false identifications be diagnosed? *Journal of Applied Psychology, 85*(4), 542.

Sobel, N. R. (1972). *Eye-miners identification: Legal and practical problems*. New York: Clark Boardman.

Sporer, S. L. (1992). Post-dicting eyewitness accuracy: Confidence, decision-times and person descriptions of choosers and non-choosers. *European Journal of Social Psychology*, *22*(2), 157-180.

Sporer, S. L. (1993). Eyewitness identification accuracy, confidence, and decision times in simultaneous and sequential lineups. *Journal of Applied Psychology, 78*(1), 22.

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.

State v. Chapple. 660 P.2d 290 (Ariz. 1983).

State v. Moon, 45 Wash App. 692, 726 P.2d 1263 (Was. Ct. App. 1986).

Steblay, N. K., Dietrich, H. L., Ryan, S. L., Raczynski, J. L., & James, K. A. (2011). Sequential lineup laps and eyewitness accuracy. *Law and Human Behavior, 35*(4), 262-274. doi:10.1007/s10979-010-9236-2

Steblay, N. K., Dysart, J. E., & Wells, G. L. (2011). Seventy-two tests of the sequential lineup superiority effect: A meta-analysis and policy discussion. *Psychology, Public Policy, and Law, 17*(1), 99-139. doi:10.1037/a0021650

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

Steblay, N., Dysart, J., Fulero, S., & Lindsay, R. C. (2001). Eyewitness accuracy rates in sequential and simultaneous lineup presentations: A meta-analytic comparison. *Law and Human Behavior, 25*(5), 459-473. doi:10.1023/a:1012888715007

Stein, J. A. (1981). Review of eyewitness testimony. *Trial Diplomacy Journal, 4*, 61–63.

Tanaka, J. W., & Farah, M. J. (1993). Parts and wholes in face recognition. *The Quarterly Journal of Experimental Psychology Section A, 46*(2), 225-245. doi:10.1080/14640749308401045

Technical Working Group for Eyewitness Evidence. (1999). *Eyewitness evidence: A guide for law enforcement*. Washington, D.C. United States Department of Justice, Office of Justice Programs.

Thompson-Cannino, J., Cotton, R., & Torneo, E. (2009). *Picking Cotton: our memoir of injustice and redemption*. New York: St. Martin's Press.

Tollestrup, P.A., Turtle, J.W., & Yuille, J.C. (1994). Actual victims and witnesses to robbery and fraud: An archival analysis. In D. Ross, D. Read, & C. Ceci (Eds.), *Adult eyewitness testimony: Current trends and developments* (pp. 144_162). New York, NY: Press Syndicate of the University of Cambridge.

Tulving, E. (1985). How many memory systems are there? *American psychologist*, *40*(4), 385.

Valentine, T. (2006). Forensic facial identifcation. In: Heaton-Armstrong, A., Shepherd, E., Gudjonsson, G. & Wolchover, D. (eds). *Witness Testimony; Psychological, Investigative and Evidential Perspectives*. Oxford: Oxford University Press.

Valentine, T., & Davis, J. P. (2015). *Forensic facial identification: theory and practice of identification from eyewitnesses, composites and CCTV*. Chichester, West Sussex, UK: Wiley Blackwell.

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., Pickering, A., & Darling, S. (2003). Characteristics of eyewitness identification that predict the outcome of real lineups. *Applied Cognitive Psychology*, *17*(8), 969-993.

Võ, M. L. H., Jacobs, A. M., Kuchinke, L., Hofmann, M., Conrad, M., Schacht, A., & Hutzler, F. (2008). The coupling of emotion and cognition in the eye: Introducing the pupil old/new effect. *Psychophysiology, 45*(1), 130-140.

Wagstaff, G. F., MacVeigh, J., Boston, R., Scott, L., Brunas-Wagstaff, J., & Cole, J. (2003). Can laboratory findings on eyewitness testimony be generalized to the real world? An archival analysis of the influence of violence, weapon presence, and age on eyewitness accuracy. *The Journal of Psychology*, *137*(1), 17-28.

Weinstein, J. (1981). Review of eyewitness testimony. *Columbia Law Review*, *81*, 441-457.

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. *Journal of Applied Social Psychology, 14*(2), 89-103. doi:10.1111/j.1559-1816.1984.tb02223.x

Wells, G. L. (2003). Murder, Extramarital Affairs, and the Issue of Probative Value. *Law and Human Behavior, 27*(6), 623-627. doi:10.1023/b:lahu.0000004890.92389.51

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

Wells, G. L., & Murray, D. M. (1983). What can psychology say about the Neil v. Biggers criteria for judging eyewitness accuracy? *Journal of Applied Psychology, 68*(3), 347.

Wells, G. L., & Olson, E. A. (2002). Eyewitness identification: Information gain from incriminating and exonerating behaviors. *Journal of Experimental Psychology: Applied, 8*(3), 155-167. doi:10.1037//1076-898x.8.3.155

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

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., 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, G. L., & Murray, D. M. (1984). Eyewitness confidence. *Eyewitness testimony: Psychological perspectives*, 155-170.

Westerman, D. L., & Larsen, J. D. (1997). Verbal-Overshadowing Effect: Evidence for a General Shift in Processing. *The American Journal of Psychology, 110*(3), 417. doi:10.2307/1423566

Wigmore, J.H. (1909). Professor Münsterberg and the psychology of testimony. *Illinois Law Review*, 3, 399–445.

Wixted, J. T. (2004). The Psychology and Neuroscience of Forgetting. *Annual Review of Psychology, 55*(1), 235-269. doi:10.1146/annurev.psych.55.090902.141555

Wixted, J. T., & Mickes, L. (2010). A continuous dual-process model of remember/know judgments. *Psychological review*, *117*(4), 1025.

Wixted, J. T., & Mickes, L. (2012). Perspectives on Psychological Science. *Perspectives on Psychological Science*, *7*(3), 275–278. http://doi.org/10.1177/1745691612442906

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

Wixted, J. T., & Stretch, V. (2004). In defense of the signal detection interpretation of remember/know judgments. *Psychonomic Bulletin & Review*, *11*(4), 616-641.

Wixted, J. T., Mickes, L., Clark, S. E., Gronlund, S. D., & Roediger, H. L. (2015). Initial eyewitness confidence reliably predicts eyewitness identification accuracy. *American Psychologist, 70*(6), 515-526. doi:10.1037/a0039510

Wixted, J. T., Mickes, L., Dunn, J. C., Clark, S. E., & Wells, W. (2016). Estimating the reliability of eyewitness identifications from police lineups. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(2), 304–309. http://doi.org/10.1073/pnas.1516814112

Wixted, J. T., Read, J. D., Lindsay, D. S., & Columbia, B. (2016). The Effect of Retention Interval on the Eyewitness Identification Confidence – Accuracy. *Journal of Applied*

Research in Memory and Cognition, *5*(2), 192–203. http://doi.org/10.1016/j.jarmac.2016.04.006

Woocher, F. D. (1977). Did your eyes deceive you? Expert psychological testimony on the unreliability of eyewitness identification. *Stanford Law Review*, *29*, 969-1030.

Woodward, A. E., Bjork, R. A., & Jongeward, R. H. (1973). Recall and recognition as a function of primary rehearsal. *Journal of Verbal Learning and Verbal Behavior, 12*(6), 608-617. doi:10.1016/s0022-5371(73)80040-4

Yerkes R.M., & Dodson J.D. (1908). The relation of strength of stimulus to rapidity of habit-formation. *Journal of Comparative Neurology and Psychology, 18,* 459–482. doi:10.1002/cne.920180503

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.

Yu, C. J., & Geiselman, R. E. (1993). Effects of Constructing Identi-Kit Composites on Photospread Identification Performance. *Criminal Justice and Behavior, 20*(3), 280-292. doi:10.1177/0093854893020003005

Yuille, J. C., & Cutshall, J. L. (1986). A Case Study of Eyewitness Memory of a Crime. *Journal of Applied Psychology*, *7*(2), 291–301.

Yuille, J. C., Cooper, B. S., Kennedy, M. A., & Herve, H. F. (2002). Weapon focus in sexual assault memories of prostitutes. *Law and Psychiatry*, *25*, 181–191.

Appendix

Chapter 3

Chapter 3: ROC Analysis - Experiment 1 and 2

The number of participants, correct ID rates, false ID rates, *pAUC* values, *D* values, and *p* values for Experiments 1 and 2.

Experiment	*N*	Correct ID rate	False ID rate	*pAUC*	*D*	*p*
			UK Lineup			
Exp 1	452	.16	.07	.008	.504	.56
Exp 2	634	.23	.08	.011		
			US Lineup			
Exp 1	509	.39	.09	.021	.042	.97
Exp 2	654	.39	.09	.021		

Note: Because the UK lineup and the US lineup did not yield significant differences in discriminability across Experiments 1 and 2, the data from both experiments were combined and presented together in Chapter 3.

Chapter 3: CAC Analysis – Experiment 1 and 2

Suspect ID accuracy across low, medium, and high levels of confidence.

Condition	*N*	Low Confidence	Medium Confidence	High Confidence
		UK Lineup		
Exp 1	452	.68 (.07)	.81 (.05)	.86 (.05)
Exp 2	634	.68 (.08)	.76 (.04)	.83 (.05)
		US Lineup		
Exp 1	509	.78 (.03)	.83 (.03)	.95 (.03)
Exp 2	654	.79 (.03)	.83 (.03)	.88 (.04)

Note: Because the UK lineup and the US lineup did not yield significant differences in reliability across Experiments 1 and 2, the data from both experiments were combined and presented together in Chapter 3. Standard error bars are shown in parenthesis.

Chapter 3: ROC Analysis - Designated Innocent Suspect

The lineup size, correct ID rates, false ID rates, and d' values for Experiments 1 and 2

Condition	*N*	Correct ID rate	False ID rate	*pAUC*	*D*	*p*
			Experiment 1 and 2			
UK Lineup	1086	.20	.15	.009	2.99	.003**
US Lineup	1163	.39	.06	.017		

Note: The most often identified filler was the designated innocent suspect. This is the other method of estimating the false ID rate (see Chapter 2). The conclusions are the same regardless of which method is used. Two ** denote a significant result at the .01 level.

Chapter 3: CAC Analysis - Designated Innocent Suspect

Suspect ID accuracy across low, medium, and high levels of confidence.

Condition	*N*	Low Confidence	Medium Confidence	High Confidence
		Experiment 1 and 2		
UK Lineup	1086	.78 (.06)	.79 (.05)	.93 (.05)
US Lineup	1163	.81 (.03)	.81 (.03)	.98 (.03)

Note: The most often identified filler was the designated innocent suspect. This is the other method of estimating the false ID rate (see Chapter 2). The conclusions are the same regardless of which method is used. Standard error bars are shown in parenthesis.

Chapter 4

Chapter 4: ROC Analysis - Designated Innocent Suspect

The number of participants, correct ID rates, false ID rates, *pAUC* values, *D* values, and *p* values for Experiments 1 and 2.

Condition	*N*	Correct ID rate	False ID rate	*pAUC*	*D*	*p*
			Experiment 3			
Verbal	358	.51	.21	.069	.27	.79
Control	359	.49	.23	.065		
			Experiment 4			
Verbal	395	.39	.15	.033	2.08	.037*
Control	375	.65	.17	.056		

Note: The most often identified filler was the designated innocent suspect. This is the other method of estimating the false ID rate (see Chapter 2). The conclusions are the same regardless of which method is used. A single * denotes a significant result.

Chapter 4: CAC Analysis - Designated Innocent Suspect

Suspect ID accuracy across low, medium, and high levels of confidence.

Condition	*N*	Low Confidence	Medium Confidence	High Confidence
			Experiment 3	
Verbal	358	.56 (.13)	.68 (.06)	.86 (.05)
Control	359	.52 (.09)	.57 (.06)	.90 (.05)
			Experiment 4	
Verbal	395	.70 (.15)	.71 (.06)	.82 (.06)
Control	375	.77 (.12)	.77 (.04)	.86 (.05)

Note: The most often identified filler was the designated innocent suspect. This is the other method of estimating the false ID rate (see Chapter 2). The conclusions are the same regardless of which method is used. Standard error bars are shown in parenthesis.

Chapter 5

Chapter 5: Perceptual Analysis – Size of the Perpetrators' Faces

The number of frames the perpetrator's face occupied, the mean size of the perpetrators face, and the standard deviation, presented in parenthesis, are displayed for each video. Individual t tests were conducted to determine whether there was a significant difference in the size of the perpetrator's face between weapon present and weapon absent videos.

	Weapon Present		Weapon Absent			
Video	# Frames	M (SD)	# Frames	M (SD)	T value	p value
Video 1	101	.17 (.07)	75	.19 (.07)	-2.7	p = .01**
Video 2	213	.22 (.17)	139	.21 (.12)	.50	p = .62
Video 3	149	.08 (.05)	114	.10 (.05)	-3.2	p < .001***
Video 4	230	.22 (.17)	223	.25 (.17)	-1.5	p = .13
Video 5	333	.15 (.09)	271	.18 (.10)	-3.8	p < .001***
Video 6	126	.14 (.02)	129	.12 (.02)	13.4	p < .001***
Video 7	172	.21 (.13)	139	.23 (.17)	-1.2	p = .25
Video 8	192	.51 (.17)	178	.41 (.14)	5.9	p < .001***

Note: The mean proportion the perpetrator's face occupied in every video is shown along with the standard deviation in parenthesis. A single * denotes a significant result, ** denote a significant result at the .01 level, and *** denote a significant result at the .001 level. Although several videos show a significant difference, it is important to note that the perpetrator's face was larger in the weapon absent video than the weapon present video in three of these videos and still no weapon focus effect was found.

Chapter 5: Perceptual Analysis – Brightness of the Perpetrators' Faces

The number of frames the perpetrator's face occupied, the mean brightness of the perpetrators face, and the standard deviation, presented in parenthesis, are displayed for each video. Individual t tests were conducted to determine whether there was a significant difference in the brightness of the perpetrator's face between weapon present and weapon absent videos.

	Weapon Present		Weapon Absent			
Video	# Frames	M (SD)	# Frames	M (SD)	T value	p value
Video 1	101	92.6 (15)	75	92.4 (11)	-.10	p = .92
Video 2	213	97.1 (27)	139	91.9 (27)	-1.6	p = .12
Video 3	149	137 (34)	114	119 (29)	-5.0	p < .001***
Video 4	230	113 (69)	223	99.6 (61)	-1.8	p = .07
Video 5	333	131 (31)	271	134 (38)	.94	p = .34
Video 6	126	112 (16)	129	121 (21)	3.8	p < .001***
Video 7	172	76 (15)	139	99 (17)	11.2	p < .001***
Video 8	192	93 (18)	178	86 (15)	-3.5	p < .001***

Note: The mean brightness of the perpetrator's face occupied in every video is shown along with the standard deviation in parenthesis. A single * denotes a significant result, ** denote a significant result at the .01 level, and *** denote a significant result at the .001 level. Although there was a significant difference in the brightness of the perpetrator's face for four videos, it is important to note that in two of these videos, the perpetrator's face was brighter in the weapon absent condition than the weapon present condition and still no significant weapon focus effect was found.