The Confidence-Accuracy Relationship for Eyewitness Reports

by

Emily Rebecca Spearing

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy in Psychology

University of Warwick, Department of Psychology

April 2022

Table of Contents

List of Tables

List of Figures

Acknowledgements

I am incredibly grateful to everyone that has supported me throughout my PhD. First and foremost, I would like to thank my supervisor, Kimberley Wade, for her tireless support and encouragement. Thank you for providing me with countless opportunities and believing in me every step of the way. I could not have wished for a better mentor or a more wonderful person to be my supervisor. You have taught me so much and I am eternally grateful for everything that you have done to help me with my research and future career.

I would also like to thank the Psychology Department at Warwick University for the financial support that has allowed me to achieve everything that I dreamed I would (and more!). Thank you to the students and academics who have made Warwick such a wonderful place to study. A special thanks to Adrian von Mühlenen, for supporting me throughout my time at Warwick University and inspiring me to undertake this PhD.

I would also like to thank the Monash Warwick Alliance for supporting my study visit to Monash University and Laura Jobson for kindly allowing me to visit her lab.

Finally, a huge thank you to my friends and family. Thanks to my parents for always believing in me and encouraging me to pursue a career that I truly care about. I would not have made it this far without your support. Thank you to Elaine and Jordan for always being there for me, and to Ross whose boundless enthusiasm about my research reminds me why I started this PhD. Thank you, Adam, for always supporting me and trying your best to understand my work. You are my best friend, and I couldn't imagine riding this rollercoaster with anyone else by my side.

Declaration

This thesis is submitted to the University of Warwick in support of my application for the degree of Doctor of Philosophy. It has been composed by myself and has not been submitted in any previous application for any degree. The work presented (including data generated and data analysis) was carried out entirely by the author.

Inclusion of Published Work

Parts of this thesis have been published by the author.

Chapters 3-5 include the following publication:

> Spearing, E. R., & Wade, K.A. (2021). Providing Eyewitness Confidence Judgments During Versus After Eyewitness Interviews Does Not Affect the Confidence-Accuracy Relationship. *Journal of Applied Research in Memory and Cognition.* Advance online publication. https://doi.org/10.1037/h0101868

Professor Kimberley Wade contributed to the planning of this research and provided feedback on the manuscript.

Chapter 6 includes the following publication:

> Spearing, E. R., & Wade, K. A. (2022). Long Retention Intervals Impair the Confidence-Accuracy Relationship for Eyewitness Recall. *Journal of Applied Research in Memory and Cognition.* Advance online publication. https://doi.org/10.1037/mac0000014

Professor Kimberley Wade contributed to the planning of this research and provided feedback on the manuscript.

Chapters 7-9 include the following publication:

> Wade, K.A., & Spearing, E.R. (under review). The Effect of Cross-Examination Style Questions on Adult Eyewitness Accuracy Depends on Question Type and Eyewitness Confidence.

Professor Kimberley Wade contributed to the planning of this research and provided feedback on the manuscript.

Abstract

Legal decision makers tend to believe that highly confident witnesses are more accurate than less confident witnesses. Furthermore, research suggests that confidence judgements can provide a useful indicator of the accuracy of witnesses' identification decisions. Yet little research has examined the relationship between eyewitness confidence and accuracy when witnesses recall details about a crime. The experiments in this thesis therefore aimed to examine the factors that influence the relationship between the accuracy of witnesses' memory reports and their confidence in those reports, and to further our understanding of how witnesses make confidence judgements.

Part One of this thesis examines several factors that may affect the relationship between confidence and accuracy when witnesses report their memories of a crime. The results of Experiments 1-3 indicate that witnesses' confidence judgements can provide a useful indicator of their memory accuracy regardless of whether they are collected immediately after each response or at the end of the memory test. Experiment 1 also indicates that the confidence-accuracy relationship can remain strong when memory performance is impaired by poor visibility, but Experiment 2 suggests that the relationship between confidence and accuracy breaks down when witnesses are exposed to misinformation. Experiment 4 explores the effect of retention interval and self-rated memory ability on the confidence-accuracy relationship. The results suggest that relatively long retention intervals can impair the confidence-accuracy relationship, but people's perception of their everyday memory ability has little impact on the confidence-accuracy relationship.

Part Two examines the confidence-accuracy relationship in the context of cross-examination. The findings suggest that the effect of cross-examination style questions on eyewitness accuracy depends on the question type and eyewitness confidence. Together, the findings presented in this thesis help to refine theories of eyewitness confidence and highlight potential issues for using confidence judgements to assess eyewitness accuracy in real cases.

Chapter 1:
Assessing the Accuracy of Eyewitness Evidence

"It is my belief that nearly any invented quotation, played with confidence, stands a good chance to deceive."

– Mark Twain

On January 6, 1990, a woman was abducted from the convenience store where she was working in Newton County, Georgia. She was dragged into the car of the assailant and sexually assaulted. Following the attack, the victim gave several statements to investigators where she described the perpetrator as a stranger. She told investigators that she had considered trying to escape from the perpetrator by running to the house of her ex-boyfriend, Ron Jacobsen, for help. Police did not initially consider Jacobsen as a suspect, until a friend of the victim's father told police that the victim was having problems with Jacobsen. The friend also told investigators that he had caught a glimpse of the assailant as he was keeping the victim company at the store. The friend later positively identified Jacobsen as the man he had seen. Despite a strong alibi and no physical evidence linking him to the crime, Jacobsen was convicted and sentenced to life in prison. He spent 30 years in prison before his conviction was overturned by DNA evidence proving his innocence.

This case, like many others, highlights the powerful influence of eyewitness testimony on the criminal justice system and demonstrates the dangers of erroneous memory evidence. Concerningly, research suggests that witness errors often contribute to wrongful convictions. Since 1989, 375 cases in the United States have been overturned due to new DNA evidence (Innocence Project, 2021). The Innocence Project estimates that at least 69% of these cases involved inaccurate eyewitness identifications. The police might come to suspect an exoneree for multiple reasons (e.g., a history of committing similar crimes), however, a recent study revealed that the most common reason is eyewitness evidence (Kenchel et al., 2020). For example, a witness may have provided a description of the perpetrator that matched the appearance of the exoneree or mentioned that they saw the exoneree in close proximity to the crime scene. Thus, it is not only mistaken identifications that lead to the wrongful conviction of innocent suspects but also

inaccuracies in witnesses' memory reports about crime details (Berkowitz et al., 2020).

Driven by the frequency of eyewitness errors, psychological researchers have investigated the factors that may impair witnesses' memories for criminal events. In these studies, psychologists typically use a mock crime methodology. Participants watch a staged crime and then, after a delay, attempt to identify the perpetrator from a lineup including several known-innocent people (i.e., "fillers"), or answer questions about the event (e.g., "What did the thief steal from the kitchen cabinet?"). In real cases, it can be difficult, if not impossible, for investigators to establish whether eyewitness reports are accurate. In lab studies, by contrast, the ground-truth is known and so the accuracy of witnesses' reports, under different experimental conditions, can be calculated.

Research using these methods has improved psychologists' understanding about the myriad factors that can impair a witness's ability to report details accurately, including factors that take effect at the time of the crime and factors that take effect at the time the witness makes an identification or provides a statement. In terms of factors that occur during the crime, research has shown that people who experience high levels of stress during the target event tend to report fewer correct details than people who experience lower levels of stress (Valentine & Mesout, 2009). Alcohol consumption (Jores et al., 2019) and the presence of a weapon (Carlson et al., 2016) can also have adverse effects on eyewitness reports.

In terms of factors that occur when a witness provides their evidence, research has shown that when investigators fail to follow best practice for collecting eyewitness reports the accuracy of the evidence can be impaired (Gabbert et al., 2015). Research shows, for example, that suggestive questioning methods (Gabbert et al., 2012; Loftus et al., 1978; Valentine & Maras, 2011), encouraging witnesses to guess (Earhart et al., 2014) and an abundance of closed questions can reduce the accuracy of witnesses' memory reports (Lamb et al., 2007). Even relatively subtle changes in the questions used to probe witnesses' memories can lead witnesses to recall details that they never experienced. One study found, for instance, that participants were more likely to falsely report seeing glass at the scene of an automobile accident when they answered the question "how fast were the cars going

when the cars *smashed* into each other?" than when the question contained the verb "hit" (Loftus & Palmer, 1974).

Research investigating how testing conditions affect eyewitness memory has informed legal decision makers on how reports should be collected from eyewitnesses. Current UK and US guidelines recommend that police interviewers refrain from using leading questions and use open questions where possible, minimising the use of closed questions (Ministry of Justice, 2011; Technical Working Group for Eyewitness Evidence, 1999). Although these guidelines may benefit the accuracy of eyewitness reports, it is important to note that there remain many factors outside the control of the criminal justice system which may serve to impair witnesses' ability to give accurate reports. Consequently, it is important that legal decision makers are informed about the factors that could undermine the accuracy of a witness's report.

Eyewitness Confidence and Legal Decision Making

One way that triers of fact may try to assess the accuracy of eyewitness reports is by monitoring the witness's confidence. Given that there is currently no guidance on how to assess eyewitness confidence during investigative interviews, police often assess confidence in a relatively informal manner by asking questions such as "are you sure that…" or "are you certain about…" (Leippe & Eisenstadt, 2007). Even implicit evaluations of eyewitness confidence could influence the course of a criminal investigation. For instance, verbal and non-verbal cues such as facial expressions, body language, hedging (e.g., "I think") and qualifying statements (e.g., "I could be wrong but…") may be interpreted as a sign of uncertainty (Lindholm et al., 2018). Indeed, any expression of low confidence may lead triers of fact to conclude that the eyewitness is inaccurate, reducing the likelihood of a guilty verdict (Bradfield & Wells, 2000; Brewer & Burke, 2002; R. C. Lindsay et al., 1981). Conversely, highly confident witnesses are likely to be believed (Grabman et al., 2021; Key et al., 2022; Slane & Dodson, 2022), regardless of other factors known to influence the accuracy of eyewitness testimony (e.g., retention interval; Cutler et al., 1988) and warnings to avoid relying on eyewitness confidence (Fox & Walters, 1986). Given the power of confidence judgements in the criminal justice system, it is important for psychologists and legal professionals to understand the

extent to which confidence judgements are predictive of the accuracy of eyewitness testimony.

Studying the Confidence-Accuracy Relationship

A large body of research spanning over 40 years has examined the relationship between eyewitness accuracy and confidence (e.g., Deffenbacher, 1980; Palmer et al., 2013; Pezdek et al., 2020). In these studies, researchers typically use a mock crime methodology and confidence judgements are commonly elicited on a numeric scale (e.g., 0-100%). Confidence judgements are sometimes taken immediately after each response (immediate-confidence judgement) and are sometimes taken at the end of the witness interview (memory test) by reminding participants of their responses (delayed-confidence judgement). The relationship between the accuracy of eyewitness reports and the confidence they express in these reports (i.e., the confidence-accuracy relationship) can then be calculated for numerous experimental conditions. The overarching goal of this research is to inform the legal system about when confidence judgements are a useful indicator of accuracy to help triers of fact distinguish between accurate and inaccurate witnesses.

The Confidence-Accuracy Correlation

Early research investigating the confidence-accuracy relationship suggested that confidence was not very informative about the accuracy of eyewitness evidence (Bothwell et al., 1987; Cutler & Penrod, 1989; Cutler et al., 1988; Deffenbacher, 1980; Leippe, 1980; Penrod & Cutler, 1995; Wells & Murray, 1984; for a review, see Wixted & Wells, 2017). The extent to which confidence discriminated between accurate and inaccurate responses was often measured by calculating the point-biserial correlation. The point-biserial correlation measures the strength of the association between confidence and binary accuracy (correct vs incorrect). A perfect correlation is obtained when all correct responses are associated with a relatively high confidence judgement and incorrect responses are associated with a lower confidence judgment.

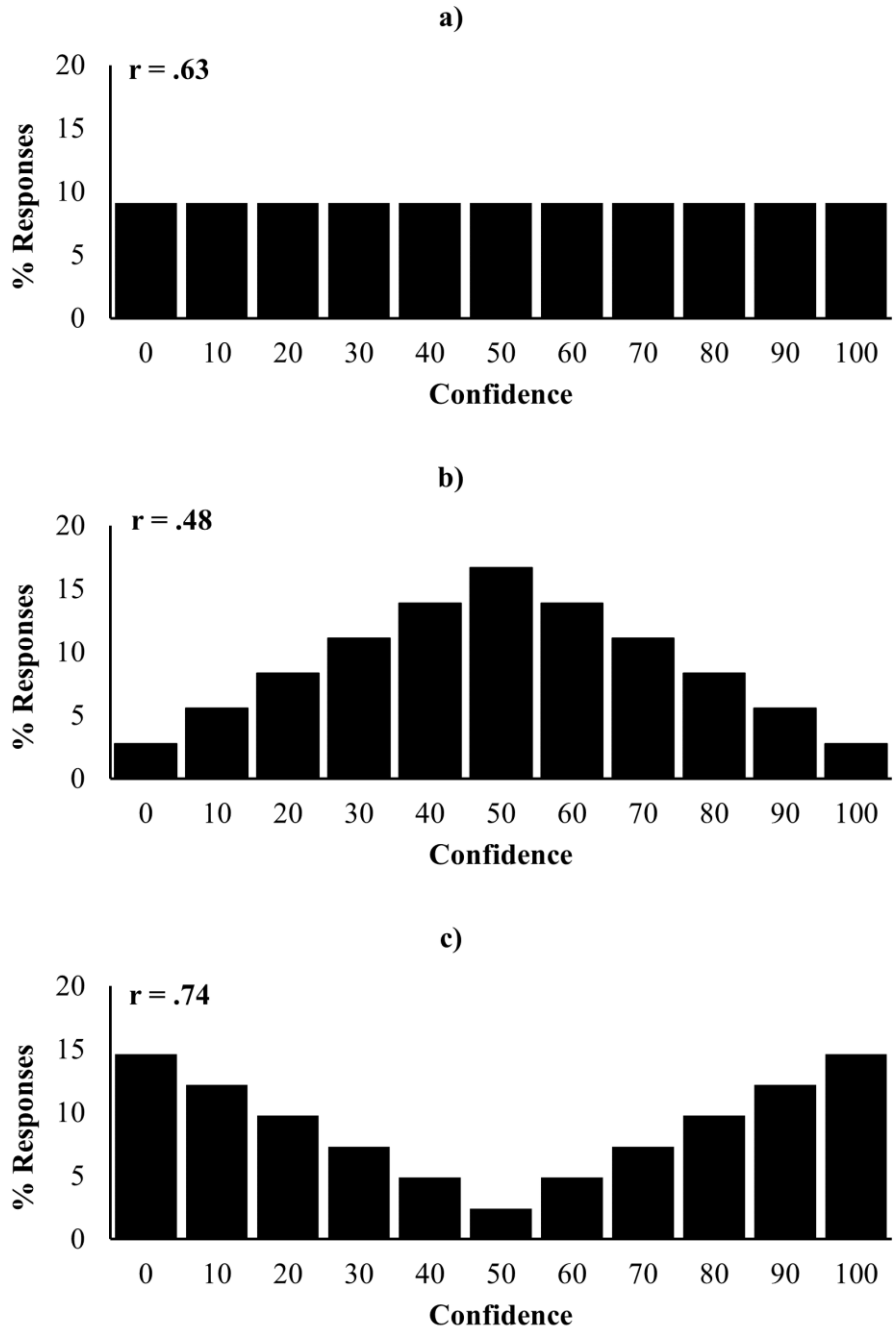
Reviews of the confidence-accuracy correlation in eyewitness research concluded that confidence statements were, at best, a modest predictor of eyewitness identification accuracy (Leippe & Eisenstadt, 2007). A meta-analysis of 31 eyewitness identification studies, for instance, found that the average correlation was

.31 (Wells & Murray, 1984). Considering these weak correlations, psychologists set out to understand the factors that influenced the strength of the confidence-accuracy relationship. Research subsequently revealed that the confidence-accuracy correlation was stronger when eyewitnesses saw the perpetrator for longer (Bothwell et al., 1987), and when analysis was limited to eyewitnesses who selected the suspect or a known-innocent foil from the lineup ("Choosers") and did not include those who said that the perpetrator was not present in the lineup ("Non-choosers"; Sporer et al., 1995).

More recently, researchers have argued that the point-biserial correlation in eyewitness studies is likely to underestimate the confidence-accuracy relationship that would be observed in real world criminal cases (Brewer, 2006; Weber & Brewer, 2003, 2004; Wixted et al., 2018). This is, at least partly, because the confidence-accuracy correlation can vary widely depending on the spread of confidence judgements (Juslin et al., 1996). Figure 1.1 shows the point-biserial correlation for three response distributions where the proportion of correct responses is perfectly matched to the confidence judgement (i.e., perfect calibration). At 40% confidence, for example, 40% of responses are correct. From looking at the correlation coefficients, one can see that the point-biserial correlation is stronger when responses are spread evenly across all levels of confidence (uniform distribution, Figure 1.1a) or when most responses are given with 0% or 100% confidence (bimodal distribution, Figure 1.1c), than when most responses are given with 40-60% confidence (unimodal distribution, Figure 1.1b). This demonstrates that the point-biserial correlation can fluctuate even when the confidence-accuracy relationship is strong, such that the proportion of correct responses is perfectly matched to the confidence judgement.

Figure 1.1

The Point-Biserial Correlation for a) a Uniform Distribution b) a Unimodal Distribution and c) a Bimodal Distribution



Note. The percentage of correct responses is equal to the percentage confidence judgement (e.g., responses made with 60% confidence are on, average, 60% accurate).

Another reason that the point-biserial correlation may underestimate the confidence-accuracy relationship in real cases is that most participants in eyewitness studies experience homogenous encoding and testing conditions, so the variation in confidence judgements is likely to be constrained (Juslin et al., 1996). This reduces the level of co-variation between accuracy and confidence, producing a correlation that is likely to underestimate the confidence-accuracy relationship that would be observed in real cases where witnessing conditions vary widely (D. S. Lindsay et al., 1998).

Indeed, the confidence-accuracy correlation tends to be stronger when witnesses experience an event under different encoding conditions. In one study, for instance, participants watched a mock crime video with good or poor encoding conditions (D. S. Lindsay et al., 2000). In the good encoding condition, participants watched a 3-minute video which showed the perpetrator from multiple perspectives and in multiple different outfits. In the poor encoding condition, by contrast, participants watched a 10-second video which showed the perpetrator from one perspective and in only one outfit. After watching the video, participants attempted to identify the perpetrator from an 8-person lineup and rated their confidence that they had identified the person from the video. Importantly, the confidence-accuracy relationship was stronger when data were collapsed across different encoding conditions (r = .55), compared to when correlations were calculated separately for good (r = .37) and poor (r = .26) encoding conditions (D. S. Lindsay et al., 2000). This finding demonstrates that the point-biserial correlation can underestimate the strength of the confidence-accuracy relationship when participants experience homogenous encoding and testing conditions.

Another criticism of the point-biserial correlation is that it does not provide the information required by legal decision makers (Brewer, 2006; Weber & Brewer, 2003; Wixted et al., 2018). While it provides a meaningful measure of resolution – that is, the extent to which correct responses are reported with higher confidence than incorrect responses – it is not informative about the probability of a witness's response being correct at a given level of confidence. For example, the point-biserial correlation does not tell lawyers the probability that the suspect is guilty given that the witness identified the suspect from the lineup with 70% confidence. Nor does it tell police the likelihood that details reported by the eyewitness, such as the type of

weapon used in the crime, are accurate. Thus, the point-biserial correlation is unlikely to help legal decision makers to discriminate between accurate and inaccurate witnesses.

Calibration Analyses

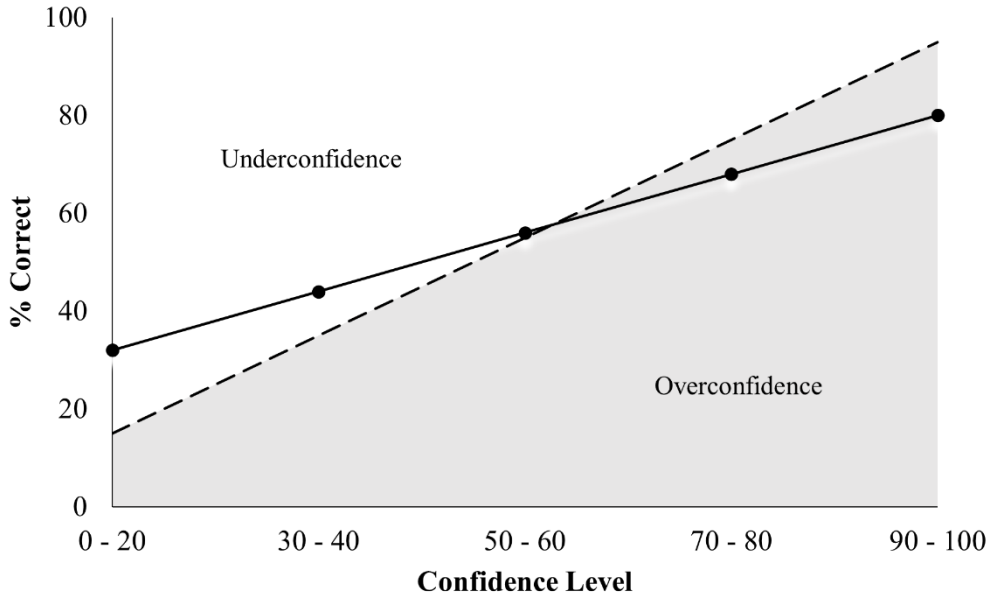
A more forensically useful measure of the confidence-accuracy relationship is calibration – that is, the extent to which the subjective likelihood (i.e., confidence) matches the objective likelihood (i.e., the proportion correct) of a correct response. Perfect calibration exists when the percentage confidence rating exactly matches the proportion of correct responses at the given level of confidence or, in other words, when responses given with 40% confidence are, on average, 40% correct. As this method is not influenced by the spread of confidence judgements, calibration can be strong even when the correlation is relatively weak. Imagine a situation where an eyewitness was 80% confident about every detail that they reported in a police interview and were, on average, 80% correct. In this situation, there is no covariation between confidence and accuracy so the correlation coefficient would be zero, despite a perfect match between the subjective and objective likelihood of a correct response (i.e., perfect calibration). While this exact scenario is unlikely to be observed in real life, any lack of variation in confidence judgements could lead to a weak correlation even when calibration is strong.

Calibration is typically assessed in two ways. First, the proportion of correct responses (i.e., accuracy) is plotted across each level of confidence and visual inspection of the resulting calibration curves shows the extent to which the observed confidence-accuracy relationship deviates from perfect calibration. Points that fall above the dashed line show underconfidence, because the percentage of correct responses is lower than would be expected at the given level of confidence if accuracy and confidence were perfectly matched. Points that fall below the dashed line show overconfidence because the proportion of correct responses is lower than would be expected for the given level of confidence. In Figure 1.2, for instance, there is underconfidence at the lowest level of confidence because the proportion of correct responses exceeds 30% whereas the given level of confidence is only 0-20%. Conversely, there is overconfidence at the highest level of confidence because the proportion of correct responses is around 80%, but the given level of confidence is

higher at 90-100%.

Figure 1.2

A Hypothetical Calibration Plot



Note. The dashed line represents perfect calibration because the proportion of correct responses is equal to the average confidence rating at each level of confidence. The points that fall below the dashed line show overconfidence because the proportion of correct responses is lower than the given level of confidence (e.g., at 90-100 confidence, only ~ 80% of responses are correct). The points that fall above the dashed line show underconfidence because the proportion of correct responses is higher than the given level of confidence.

Second, Calibration (*C*) and Over/Underconfidence (*OU*) statistics are calculated (Lichtenstein et al., 1977). The *C* statistic measures the total amount of deviation from perfect calibration and ranges from 0 (*perfect calibration*) to 1. The over/underconfidence (*OU*) statistic measures the extent to which confidence judgements generally under- or overestimate response accuracy and ranges from -1 (*complete underconfidence*) to 1 (*complete overconfidence*). Whereas calibration curves provide information about the amount of over- or underconfidence at a given level of confidence within an entire sample, *C* and *OU* statistics are calculated for

each participant. Thus, these statistics help psychologists to examine the extent to which the confidence-accuracy relationship differs between individuals.

An advantage of the calibration approach is that it provides the information required by the legal system. Specifically, inspection of calibration curves can reveal the likely accuracy of a response at a given level of confidence. Looking at the hypothetical calibration curve in Figure 1.2, one can see that the probability of a correct response increases from ~30% at the lowest level of confidence to ~80% at the highest level of confidence. If this pattern is observed in forensic contexts, then confidence could provide useful information for triers of fact for assessing the accuracy of eyewitness reports.

Calibration studies have changed memory scientists' view of the confidence-accuracy relationship. Whereas low correlations previously instilled pessimism about the strength of the confidence-accuracy relationship, calibration research has shown that confidence can be a reliable indicator of identification accuracy. But there are some crucial caveats. Research suggests that confidence is only a reliable predictor of memory accuracy when confidence judgements are taken under relatively "pristine conditions" (Wells et al., 2020; Wixted et al., 2018; Wixted & Wells, 2017). What are these pristine conditions? First, the confidence judgement must be taken immediately after the witness makes an identification decision and the witness should only be tested once (Brewer, 2006). Second, the suspect should not unfairly stand out in the lineup (e.g., due to some distinctive feature; Colloff et al., 2016). Third, the witness should be warned that the perpetrator may not be in the lineup (Quinlivan et al., 2012). Fourth, the witness should not be exposed to feedback that supports their identification (e.g., a police officer says to the witness "Great, you got our suspect!"; Bradfield et al., 2002; Douglass & Steblay, 2006; Semmler et al., 2004). When all of these conditions are met, research suggests that highly confident eyewitnesses are likely to be highly accurate.

Yet some memory experts continue to urge caution about using confidence to assess eyewitness accuracy (Berkowitz et al., 2020; Sauer et al., 2019; Wade et al., 2018). They warn that we do not fully understand the factors that might impair the confidence-accuracy relationship in real cases, and it is difficult, if not impossible, to ensure that witnesses' memories are not contaminated before they make an

identification or provide a memory report. Furthermore, most calibration studies have investigated the strength of the confidence-accuracy relationship in the context of eyewitness identification. As a result, relatively little is known about eyewitness accuracy and confidence when witnesses are asked to recall details about a crime, such as in a police interview. There are many unanswered questions about how confidence judgements should be taken during police interviews, and the factors that harm the confidence-accuracy relationship for eyewitness recall. When should confidence judgements be taken? And what type of question is harmful to eyewitness calibration? Investigating these issues will help psychologists to understand when confidence judgements are useful for estimating the accuracy of eyewitness reports and to develop a better theoretical understanding of the basis of eyewitness confidence.

Summary

In sum, eyewitness evidence is one of the leading causes of wrongful conviction, and highly confident witnesses are likely to be believed regardless of whether they are accurate. Whereas early research suggested that confidence was not very informative about the accuracy of eyewitness evidence, recent work shows that the confidence-accuracy relationship can be strong when calibration is calculated. Still, relatively little is known about the confidence-accuracy relationship when witnesses report details about the witnessed event. The next chapter outlines an influential theory of eyewitness confidence and existing work on the confidence-accuracy relationship for eyewitness reports.

Chapter 2:
Eyewitness Confidence: Theory and Research

Introduction

Psychologists have been developing and refining theories of how people make confidence judgements for over 40 years (Griffin & Tversky, 1992; Koriat, 1997; Koriat et al., 1980; Leippe et al., 2009). These theories aim to explain how people assess the accuracy of their memories and have been used to guide research not only in psychology and law, but also in other applied areas (e.g., education). The current chapter focuses on one influential account of how people make confidence judgements, that is commonly used in the eyewitness memory literature: Koriat's (1997) cue-utilisation theory.

Koriat's (1997) Cue-Utilisation Theory

According to cue-utilisation theory, people use numerous experience-based and theory-based cues to assess the likely accuracy of their memories and adjust their confidence judgements accordingly (Koriat, 1997). Experience-based cues refer to any information obtained during the process of remembering, such as the speed or ease with which information is recalled from memory (Koriat & Levy-Sadot, 1999; Koriat et al., 2008). Theory-based cues, by contrast, refer to people's meta-memorial beliefs about how memory works and their ability to remember information accurately. For example, people generally believe that it is easier to recall information after a short retention interval than after a long retention interval (Cormia et al., 2020). A key tenet of the theory is that the strength of the confidence-accuracy relationship depends on whether these cues are objectively related to memory performance (Koriat, 1997). If people rely on cues that are not informative about their memory accuracy, then they are likely to conclude that their memory is more or less accurate than it really is, impairing the confidence-accuracy relationship. Throughout this thesis, cue-utilisation theory is used to predict when confidence judgements are likely to provide a useful indicator of accuracy, and to explain why the confidence-accuracy relationship sometimes breaks down.

Experience-Based Cues

According to cue-utilisation theory, the strength of the confidence-accuracy relationship will, at least partly, depend on the extent to which experience-based cues such as ease of retrieval are objectively related to memory accuracy. For example, if information that comes to mind relatively quickly is more accurate than information that comes to mind relatively slowly then monitoring the speed with which information comes to mind may help people to reach a reasonable estimate of their memory accuracy. Put simply, confidence judgements are likely to provide a useful indicator of memory accuracy when people correctly use experience-based cues that are objectively related to their memory accuracy. Confidence judgements are likely to be misleading about memory accuracy, however, if people incorrectly use experience-based cues that are not related to their memory performance.

A growing body of research supports the idea that people use experience-based cues to guide their confidence judgements (Gustafsson et al., 2019, 2021; Lindholm et al., 2018; Robinson et al., 1997). Studies investigating the influence of experience-based cues on confidence judgements have often used response times as an indicator of retrieval fluency – that is, the ease with which information comes to mind. Measuring response times allows researchers to investigate the situations in which retrieval fluency is likely to be informative about memory accuracy, and when it is not. Decades of research shows that retrieval fluency usually provides a useful indicator of memory accuracy: the more quickly a response is made, the more likely it is that the response is accurate. For instance, when witnesses identify a perpetrator from a lineup, they are more likely to have identified the perpetrator correctly and with high confidence if they made the identification quickly rather than slowly (Dodson & Dobolyi, 2016; Sauerland & Sporer, 2007; Seale-Carlisle et al., 2021; Sporer, 1992). Likewise, when witnesses answer questions about a crime, responses given relatively quickly tend to be more accurate and reported with higher confidence than responses given relatively slowly (Robinson et al., 1997). These findings support the notion that people use retrieval fluency to judge the likely accuracy of their memories and adjust their confidence judgements accordingly.

Linguistic cues also provide insight into the amount of cognitive effort required to retrieve a piece of information from memory. In one study, participants

watched a videotape of a woman being abducted by two men (Lindholm et al., 2018). After watching the mock crime event, participants completed a free recall task and answered 12 cued recall questions about the perpetrator (e.g., "Describe the clothing of the man who waited outside the victim's house"). Their responses were coded for accuracy and several linguistic and paralinguistic cues, including delays (i.e., pauses before or during a response), fillers (e.g., "um", "uh"), and hedges (e.g., "I guess"). Responses containing a greater number of hedges and fillers, and fewer words, were generally assigned lower confidence judgements than responses containing fewer cues of uncertainty (Lindholm et al., 2018). Furthermore, hedges fully mediated the confidence-accuracy correlation. Together, these results suggest that retrieval fluency can often provide a useful indicator of memory accuracy, and people often use this information to make their confidence judgements.

Although experience-based cues usually provide a good indicator of memory accuracy, they can sometimes lead witnesses to adjust their confidence inappropriately, impairing the confidence-accuracy relationship. For instance, when people were exposed to correct answers and related but incorrect answers to a subsequent general knowledge test, they recalled these answers more quickly and with higher confidence than correct answers that they had not been shown previously (Kelley & Lindsay, 1993). Put simply, prior exposure to an answer increased the ease with which that answer came to mind, and this increased retrieval fluency was interpreted as a sign that the answer was accurate, inflating confidence judgements.

This increased-fluency mechanism may also explain why witnesses sometimes come to report misinformation with high confidence. In the typical misinformation experiment, participants witness a mock crime and, after a delay, are exposed to misinformation. This post-event information is typically presented in a narrative or misleading questions or provided by a co-witness. Finally, participants answer questions about the crime (Gabbert et al., 2004; Hope et al., 2008; Ito et al., 2019; Luna & Martín-Luengo, 2012; Paterson & Kemp, 2006). Studies using this methodology show that people often come to report information that they have never witnessed with high confidence (Bonham & González-Vallejo, 2009; Flowe et al., 2019). Why? One explanation is that, because the misinformation is received after the crime, it often comes to mind more quickly than information from the original event. Furthermore, research suggests that people do not spontaneously monitor the

source of their memories, so they often fail to realise when information did not come from the original event (Horry et al., 2014; D. S. Lindsay & Johnson, 1989). According to cue-utilisation theory, people interpret the increased retrieval fluency as a sign that this post-event information is accurate and therefore report it with high confidence, impairing the confidence-accuracy relationship (Koriat, 1997).

Theory-Based Cues

According to Koriat's (1997) cue-utilisation theory, people also rely on their meta-memorial beliefs about how memory works and their ability to remember information accurately to make confidence judgements. For example, people may believe that it will be easier to recognise someone that they saw close-up and therefore identify them with higher confidence than someone they saw from further away. If people use theory-based cues to assess their memory accuracy, then the strength of the confidence-accuracy relationship should rely on the accuracy of people's meta-memorial beliefs. If witnesses realise that their memory performance has been affected by a given factor then they may adjust their confidence ratings accordingly, maintaining the confidence-accuracy relationship. If, however, witnesses fail to realise that their memory has been compromised then they may fail to adjust their confidence to reflect their memory accuracy, impairing the confidence-accuracy relationship.

How accurate, then, are people's meta-memorial beliefs about memory performance? Surveys show that laypeople have accurate beliefs about how some factors influence memory performance but have inaccurate beliefs about other factors (Ost et al., 2017). A recent survey, for instance, found that people had accurate beliefs about some variables including the effect of lighting and alcohol intoxication on memory performance (Cormia et al., 2020). However, people had inaccurate beliefs about other variables including the effect of marijuana intoxication on memory performance and the effect of race on eyewitness identification. Based on Koriat's (1997) cue-utilisation theory, it might be expected that the confidence-accuracy relationship will be relatively strong when people are exposed to factors that they have accurate beliefs about (i.e., alcohol intoxication), but not when they are exposed to factors that they have inaccurate beliefs about (i.e., marijuana intoxication).

Although there is little research directly investigating the role of theory-based cues, a growing body of evidence is consistent with this interpretation. The confidence-accuracy relationship is maintained, for example, when memory performance is impaired due to alcohol consumption (Flowe et al., 2019). According to cue-utilisation theory, confidence judgements remain a useful indicator of accuracy because people know that alcohol consumption impairs memory performance and lower their confidence judgements to compensate for their lower accuracy, maintaining the confidence-accuracy relationship (Koriat, 1997).

Conversely, the confidence-accuracy relationship is impaired when witnesses are intoxicated with marijuana (Pezdek et al., 2020). In one study, participants viewed 24 white faces while intoxicated with marijuana, or not under the influence of any substance. Immediately after viewing the faces, participants saw 48 faces and attempted to identify which faces they had seen previously and rated their confidence in their responses. Unintoxicated participants were significantly better at discriminating between new and old faces than participants intoxicated with marijuana. Furthermore, marijuana intoxication impaired the confidence-accuracy relationship: unintoxicated participants were 85% accurate at the highest level of confidence, whereas intoxicated participants were only 68% accurate (Pezdek et al., 2020). Put simply, people who were intoxicated with marijuana failed to adjust their confidence to compensate for their lower accuracy. These findings are consistent with cue-utilisation theory, which suggests that the confidence-accuracy relationship will be impaired if people do not realise that a given factor impairs memory performance and thus fail to adjust their confidence ratings to reflect their lower accuracy (Koriat, 1997).

In sum, cue-utilisation theory suggests that people base their confidence judgements on numerous experience-based cues and theory-based cues (Koriat, 1997). Consistent with this theory, research suggests that people tend to be more confident in information that comes to mind relatively quickly and that takes relatively little cognitive effort to recall. Moreover, people seem to be better at judging the accuracy of their memories when their memories are affected by factors that they have accurate beliefs about, than factors that they have inaccurate beliefs about. Guided by this theory, memory researchers can make predictions about how different experimental conditions will affect witnesses' confidence in their reports,

and the conditions in which the confidence-accuracy relationship is likely to break down.

The Confidence-Accuracy Relationship for Eyewitness Reports

Although many studies have examined the confidence-accuracy relationship when witnesses make identifications from lineups (e.g., Palmer et al., 2013; Sauer et al., 2010), studies have rarely examined the confidence-accuracy relationship when witnesses report details about a crime. The few studies that have examined the confidence-accuracy relationship for eyewitness reports suggest that confidence can provide a good indicator of accuracy in some conditions (Flowe et al., 2019; Horry et al., 2014; Odinot et al., 2009; Robinson & Johnson, 1996; Saraiva et al., 2020; Sauer & Hope, 2016). In these studies, participants typically watch a mock crime video and, after a delay, answer cued recall questions about details in the video (e.g., the colour of the perpetrator's clothing) or report everything that they can remember about the video. In one study, participants answered cued recall questions about a simulated bank robbery, and immediately rated their confidence in each response from 0% to 100% (Luna & Martín-Luengo, 2012). Responses made with 100% confidence were over 20% more likely to be accurate than responses made with 80% confidence. When participants freely report details about a crime, their high confidence responses tend to be even more accurate than when they answer cued recall questions. Indeed, another study found that details reported with 90-100% confidence were correct over 95% of the time (Saraiva et al., 2020). Together, these findings suggest that confidence judgements may provide useful information about the accuracy of witnesses' reports, at least when memory is relatively strong.

There is also evidence to suggest that the confidence-accuracy relationship can remain strong when witnesses experience suboptimal encoding conditions which impair their memory performance. One study showed that dividing attention during encoding impaired the accuracy of eyewitness reports, but not the confidence-accuracy relationship (Sauer & Hope, 2016). Participants encoded 4 images depicting different scenarios and answered cued recall questions about each image. In the divided-attention condition, participants simultaneously completed a secondary task during encoding, which involved pressing a key when they heard an odd number in a sequence of randomly ordered numbers. Participants in the full-

attention condition did not complete a secondary task. When participants' attention was divided between encoding and a secondary task, participants provided less accurate reports and fewer specific details than when participants paid full attention to the images (Sauer & Hope, 2016). Nonetheless, divided attention did not affect witnesses' ability to monitor the accuracy of their memories. These findings suggest that witnesses sometimes reduce their confidence judgements to reflect their lower memory accuracy when their memory is impaired, maintaining the confidence-accuracy relationship.

There are some circumstances, however, in which witnesses' fail to adjust their confidence judgements to reflect their lower accuracy. For example, witnesses who are exposed to misinformation about a crime often come to report this information and sometimes do so with high confidence, producing overconfidence (Bonham & González-Vallejo, 2009; Flowe et al., 2019; Horry et al., 2014). Furthermore, factors in the testing environment can lead witnesses to believe that their memory reports are more or less accurate than they really are. Confidence judgements can be inflated, for instance, when witnesses receive positive feedback about their memory ability (e.g., "Your eyewitness testimony tends to be extremely accurate"; Iida et al., 2021) or when witnesses answer easy questions before seemingly more difficult questions (Michael & Garry, 2016).

Another factor that can affect the confidence-accuracy relationship is the type of information that witnesses report. Research shows, for example, that confidence tends to provide a better indicator of accuracy for central information than peripheral information (Roberts & Higham, 2002; Sarwar et al., 2014). The confidence-accuracy relationship can also vary with the specificity of witnesses' reports. Recent research suggests that confidence is more informative about accuracy for specific, "fine-grain" details (e.g., "navy blue") than more general, "coarse-grain" details (e.g., "dark"; Goldsmith et al., 2002; Sauer & Hope, 2016; Vredeveldt & Sauer, 2015; Weber & Brewer, 2008). In one study examining the confidence-accuracy relationship for general and specific details, participants watched a video depicting a violent encounter then completed a 5-minute filler task (Vredeveldt & Sauer, 2015). Participants than answered 20 cued recall questions about the video, and immediately rated their confidence for each response from 0% (*not confident at all*) to 100% (*extremely confident*). Compared to general details, specific details were

reported with higher confidence and showed significantly less underconfidence. Thus, confidence was more informative about the accuracy of specific details than general details.

In sum, existing work on the confidence-accuracy relationship for eyewitness reports suggests that confidence judgements can provide a useful indicator of accuracy in some conditions. Importantly, the confidence-accuracy relationship can be strong even when memory is impaired. Thus, decreases in accuracy are not always mirrored by a reduction in the confidence-accuracy relationship. However, factors such as misinformation and feedback can impair witnesses' ability to monitor the accuracy of their memories, which can lead the confidence-accuracy relationship to break down. Even when witnesses are interviewed appropriately without suggestive influences, confidence judgements may provide more useful information about central and specific information than general and peripheral information.

Thesis Aims and Outline

This thesis consists of two parts and aims to advance our understanding of the relationship between accuracy and confidence when witnesses report details about a crime. Although confidence is often used to gauge the accuracy of eyewitness testimony, there is relatively little research examining whether confidence judgements are informative about the accuracy of witnesses' memory reports. Moreover, few studies have tested the predictions born out of theoretical models of confidence judgements (i.e., cue-utilisation theory) in the context of eyewitness memory reports.

Given the powerful influence of eyewitness confidence on legal decision making, it is important that a) investigators know when confidence judgements are a reliable indicator of memory accuracy and are aware of the factors that influence eyewitness confidence, and b) that confidence judgements are collected in a way that maximises their informativeness about eyewitness accuracy. Therefore, the research presented in Part One examines several factors that may affect the relationship between eyewitness confidence and accuracy when witnesses report details about a crime. Part Two builds on Part One by exploring the relationship between confidence and accuracy in a new context: cross-examination.

The overarching aim of this thesis is to investigate the factors that influence the relationship between the accuracy of eyewitnesses' memory reports and the confidence that witnesses express in their reports. From an applied perspective, this research may help to produce research-led guidelines and policy on how best to elicit confidence judgements, and the conditions in which confidence judgements are likely to be informative about the accuracy of eyewitness reports. From a theoretical perspective, investigating the factors that underlie eyewitness confidence judgements will improve our theoretical understanding of how witnesses make confidence judgements and further refine theories on eyewitness confidence and accuracy.

This thesis outlines a series of 6 experiments exploring how several different factors affect the confidence-accuracy relationship for witness reports. All of the experiments were pre-registered, and the data are freely available on the Open Science Framework (OSF). These experiments were guided by Koriat's (1997) cue-utilisation theory, as well as current practical concerns about using confidence judgements to assess the accuracy of eyewitness testimony.

Part One includes Chapters 3 to 6. Chapter 3 begins to examine whether the timing of witnesses' confidence judgements affects the relationship between the accuracy of their memory reports and their confidence in those reports. This chapter proposes that confidence judgements may provide a better indicator of accuracy when they are collected immediately after each response, than when they are taken after witnesses' have reported everything that they can remember.

In Chapters 4 and 5, Experiments 1-3 test the prediction made in Chapter 3. Each experiment used a mock crime paradigm and participants rated their confidence in their responses to the memory test, either immediately after each response or at the end of the memory test. Chapter 4 also examines whether people can adjust their confidence judgements to reflect the accuracy of their reports when their memory performance is impaired by (a) poor visibility (Experiment 1) and (b) exposure to misinformation (Experiment 2). In Chapter 5, Experiment 3 investigates whether the timing of confidence judgements affects the number of details that witnesses report about the crime. Together, these chapters provide more information about when confidence judgements could be collected, and the circumstances in which

confidence judgements are likely to be informative about the accuracy of eyewitness reports.

In Chapter 6, Experiment 4 investigates how the retention interval between viewing a crime and providing a statement (i.e., the retention interval) affects people's ability to monitor the accuracy of their responses. Experiment 4 also builds on our understanding of the role of individual differences by examining whether the effect of retention interval on eyewitness confidence depends on how people feel about their memory ability. To this end, participants answered questions about their general memory ability and watched a mock crime video. After watching the video, participants either proceeded to the memory test immediately, or after a delay of 1 week or 1 month.

Part Two includes Chapters 7 to 9. Chapter 7 begins to explore whether cross-examination style questions enhance eyewitness accuracy and proposes that the effect of cross-examination style questions may depend on the type of question, and how confident witnesses' feel about their memory for the original event.

Experiment 5, reported in Chapter 8, tests the claims made in Chapter 7 and builds on Part One by assessing the confidence-accuracy relationship in the context of cross-examination. Participants answered two memory tests 3-5 days apart and rated their confidence in their responses. During the second (cross-examination) memory test, some participants answered a mixture of cross-examination style and simple questions, whereas other participants only answered simple questions like those in Experiments 1-4.

The aim of Chapter 9 was to replicate the results of Experiment 5 when participants answered fewer cross-examination style questions. In Experiment 6, all participants answered a mixture of simple and cross-examination style questions in the second memory test, and the number of cross-examination style questions was reduced from Experiment 5. Together, Experiments 5 and 6 are the first to systematically examine how witnesses' confidence in their memory and their ability to monitor the accuracy of their reports affects their performance during cross-examination.

Finally, Chapter 10 brings together the findings of Experiments 1-6 in the General Discussion. The practical implications for legal decision makers (e.g.,

lawyers, jurors) and policy makers are outlined, as well as the theoretical implications for Koriat's (1997) cue-utilisation theory which hypothesises about how people judge the accuracy of their memory reports.

Part One

Chapter 3:
The Timing of Confidence Judgements in Police Interviews

As outlined in Chapter 1, decades of research have explored the extent to which witnesses' confidence judgements are informative of their memory accuracy. Early research focused on whether a witness's confidence in their identification decision from a lineup was a reliable indicator of the accuracy of that decision and suggested that confidence and accuracy were weakly related (Deffenbacher, 1980; Penrod & Cutler, 1995; Wells & Murray, 1984). In line with these findings, analyses of real-world DNA exoneration cases have revealed that erroneous eyewitness testimony has long been the leading cause of wrongful convictions (Huff, 1987; Scheck et al., 2000), and that erroneous in-court testimony is frequently accompanied by inflated confidence (Garrett, 2011). A new wave of research, however, has shown that eyewitness confidence—at least in some contexts—is a better predictor of memory accuracy than originally thought (Brewer & Wells, 2006; Mickes, 2015; Wixted & Wells, 2017). These latest findings suggest that high confidence responses are highly likely to be accurate when witnesses' memories are not contaminated by misinformation or improper procedures.

Yet, some memory experts continue to warn about the risks of overstating the value of witness confidence in criminal investigations. They have posited that we do not have enough data to fully understand the factors that influence the confidence-accuracy relationship in real cases (Berkowitz & Frenda, 2018; Berkowitz et al., 2020; Sauer et al., 2019; Wade et al., 2018). Furthermore, most of what we know about eyewitness confidence and accuracy is based on eyewitness identification research—studies in which people make a single decision in an attempt to identify a perpetrator from a lineup (Colloff et al., 2017; Palmer et al., 2013). Relatively few studies have explored the confidence-accuracy relationship in witness interviews, where mock-witnesses are asked to recall crucial details, such as a perpetrator's features, clothing, and actions. In Experiments 1-3, we investigated the confidence-accuracy relationship when witnesses report details about a crime and gathered data on a factor that might impair this association: The timing of confidence ratings.

As noted in Chapter 2, the few studies that have explored the confidence-accuracy relationship in witness interviews suggest that confidence may be a good proxy for accuracy in some contexts (Odinot et al., 2013; Weber & Brewer, 2008). For instance, a recent study found that confidence was strongly related to accuracy, regardless of whether details were elicited through open-ended interview instructions (i.e., "Write down everything you can remember") or forced report cued recall questions (e.g., "What was the hair colour of the customer?"; Brewer et al., 2018). Other studies have shown that the confidence-accuracy relationship remains strong even when memory is weak because people tend to adjust their confidence when they recognise that their memory has been compromised (Koriat, 1997). For example, participants have appropriately lowered their confidence to compensate for their lower accuracy after their attention was divided between encoding and a secondary task (Sauer & Hope, 2016). Taken together, this research suggests that confidence can be informative about the accuracy of eyewitness reports at least in some circumstances.

There are circumstances, however, in which the confidence-accuracy relationship breaks down. A growing number of studies reveal that when people are unknowingly exposed to misinformation, they often report this information with high confidence (Flowe et al., 2019; Hope et al., 2008; Wright et al., 2000). Confidence can also be inflated by factors that do not affect memory accuracy, such as when people receive positive feedback about their memory ability (e.g., "you're spot on"; Iida et al., 2020) or answer easy questions before difficult questions in a memory test (Michael & Garry, 2016). These studies suggest that the context in which witnesses' confidence and memory reports are gathered are vital to preserving a strong confidence-accuracy relationship.

One key testing condition, however, that has been largely ignored in the eyewitness literature, is the timing of witnesses' confidence judgements. A review of published research shows that some studies have collected confidence immediately after participants reported each detail (*immediate-confidence* judgements; e.g., Dodson & Krueger, 2006; Paulo et al., 2016), whereas others have collected confidence only after participants reported everything they could remember (*delayed-confidence* judgements; e.g., Evans & Fisher, 2011; Roberts & Higham, 2002). Given that the methodologies differed substantially across these studies—

including the nature of the stimuli, the retention intervals, the interview procedures, and testing formats—it is difficult to ascertain how the timing of confidence judgements might have influenced the confidence-accuracy relationship.

To date, only one published study has systematically compared immediate- and delayed-confidence judgements. In a standard witness memory experiment, participants watched a mock crime video, then, after a 7-minute delay, answered questions about the event and rated their confidence in their responses either immediately after each question or at the end of the memory test (Robinson & Johnson, 1996). The data revealed no differences in confidence ratings between the immediate and delayed confidence groups for both recall and recognition memory. Moreover, there was no positive relationship between confidence and accuracy: Neither immediate- nor delayed-confidence judgements reliably distinguished between correct and incorrect responses (known as *resolution*). In this study, confidence-accuracy calibration was never examined, yet it is needed to assess the level of under- and overconfidence at each level of confidence (Brewer & Wells, 2006). Furthermore, the encoding conditions in this study were homogenous—all participants viewed the same perpetrator and crime under the same viewing conditions—so the resulting confidence-accuracy correlation almost certainly underestimates the correlation that would be observed in real-world witnessing situations (D. S. Lindsay et al., 1998, 2000).

There are at least two reasons to predict that the timing of confidence judgements might influence the confidence-accuracy relationship. Previous work suggests witnesses' confidence could become inflated when their confidence judgements are delayed. Many studies show that repeated exposure to information can increase the ease with which people process that information, which in turn increases the subjective impression that the information is accurate (Alter & Oppenheimer, 2009; Kelley & Lindsay, 1993; Unkelbach & Stahl, 2009). People report, for instance, being more confident they have visited a specific location, like a university campus, after viewing a photo of that location twice rather than only once (Brown & Marsh, 2008). People are also more confident that an eyewitness's erroneous claims are true after reading those claims three times rather than once (Foster et al., 2012). One important distinction between immediate- and delayed-confidence judgements is that when confidence judgements are made immediately,

witnesses only consider their responses at the point of retrieval. But when confidence judgements are delayed, witnesses consider their responses at the point of retrieval and then again, after a delay, when they are asked to provide confidence judgements for each detail they have recalled. In short, when confidence judgements are delayed, witnesses are repeatedly exposed to their responses, which could enhance processing fluency and lead to overconfidence. If so, we might expect immediate-confidence participants to show stronger confidence-accuracy calibration than delayed-confidence participants. Unpublished lineup research supports this prediction, showing that confidence judgements taken immediately after an identification produced stronger calibration than confidence judgements taken after a short 5-minute delay (Brewer et al., 2005).

Determining whether the timing of confidence ratings affects the confidence-accuracy relationship is both practically and theoretically important. On the practical side, triers of fact often rely on a witness's confidence to judge the accuracy of their testimony, even when it's obvious that the witness has been exposed to factors known to reduce memory accuracy (e.g., the presence of a weapon; Bradfield & Wells, 2000; Cutler et al., 1988; Fox & Walters, 1986). As such it is vital to understand the conditions in which the confidence-accuracy relationship breaks down. Also, if the confidence-accuracy relationship is stronger when confidence judgements are collected during rather than after memory retrieval, then we may reveal a simple investigative procedure that can help interviewers to enhance the confidence-accuracy relationship. On the theoretical side, examining how confidence timing influences the confidence-accuracy relationship may advance understanding of how witnesses determine the accuracy of their reports. Although research suggests that confidence can provide some predictive value about the accuracy of eyewitness statements, research examining the basis of these confidence judgements is limited.

In Experiments 1-3, we examined whether the timing of confidence judgements affects the confidence-accuracy relationship in eyewitness interviews. In each experiment, participants watched a mock crime video, and after a delay completed a memory test and rated their confidence in their responses. Some participants rated their confidence immediately after providing each response, and others rated their confidence at the end of the memory test. Each experiment was preregistered; the numeric data for all experiments, and the corresponding R code are

available on the Open Science Framework: https://osf.io/mp3r8/ for Experiment 1, https://osf.io/gqkyp/ for Experiment 2, and https://osf.io/dbmnc/ for Experiment 3.

Chapter 4:
Does the Timing of a Witness's Confidence Judgements Affect the Confidence-Accuracy Relationship?

Experiment 1

Method

Participants & Design

It is important to create variability in encoding conditions when trying to detect reliable and generalizable effects in witness memory research (Brewer et al., 2010; D. S. Lindsay et al., 1998). Accordingly, we manipulated the crime event (i.e., the type of crime and the perpetrator viewed) and the visibility of the crime event (i.e., day versus night visibility) so that encoding conditions varied on several dimensions. We used a 2 (Event: car theft, mugging) x 2 (Visibility: day, night) x 2 (Confidence timing: immediate, delayed) between-participants design. Calibration studies typically employ large samples, with more than 200 observations per condition, to achieve stable estimates (Brewer & Wells, 2006). Based on these studies, and to attain stable calibration estimates, we aimed to collect 18 observations from at least 320 people, producing ~720 observations in each of the eight conditions.

In total, we recruited 380 adults from Canada, the United Kingdom, and the United States through Amazon's Mechanical Turk (MTurk) using the CloudResearch platform (Litman et al., 2017). Participants received $1.25 for completing the experiment. We excluded those who answered an attention check question incorrectly (n = 13), experienced technical difficulties (n = 9) or reported that they failed to comply with any of the criteria outlined in the experiment (e.g., if they watched the video more than once, n = 18). The final sample consisted of 340 participants (194 male, 143 female, 3 undisclosed, M = 37.29 years, SD = 11.55, range = 20-71), producing 6,120 observations in total. There were 41-48 participants in each of the 8 between-participant groups. The Department of Psychology Research Ethics Committee at the University of Warwick approved this research.

Figure 4.1

Screenshots of the Car Theft (Left) and Mugging (Right) Videos with Day (Top) and Night Visibility (Bottom)

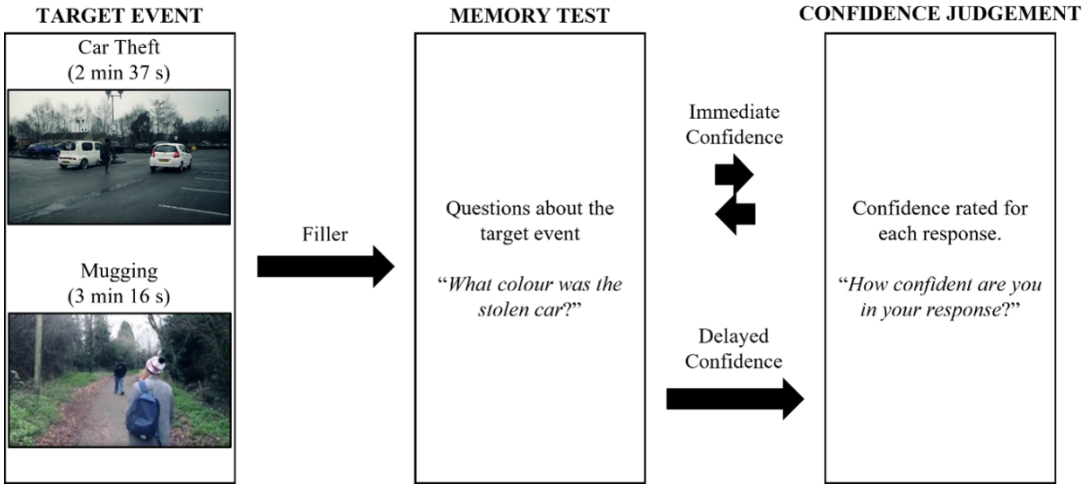
Materials

Videos. We created two mock crime videos for the experiment, a car theft and a mugging scenario. In the car theft scenario (2 min 37 s), a female thief walks around a supermarket car park and peeks into a parked car. She notices the female victim leaving her car and goes to steal it. The victim sees the thief driving away and chases after the car. In the mugging scenario (3 min 16 s), a female victim exchanges phone numbers with a male friend. She puts the phone into her bag and a male thief instructs her to give him the bag. When she refuses, he wrestles the bag from her and runs away. Using Adobe After Effects®, we digitally altered the two original videos so they resembled a night scene, producing a total of four videos (see Figure 4.1).

Memory tests. We created a memory test for each crime scenario that contained 18 4-alternative forced choice (4AFC) questions (see Appendix A for the full memory test). Questions pertained to people, actions, objects, and locations in the video (e.g., “What was the colour of the stolen bag?”). Each question was presented with a correct answer and three incorrect alternatives (e.g., correct answer: blue, incorrect alternatives: grey, black, brown).

Figure 4.2

The General Procedure for Experiments 1-3



Procedure

A general overview of the procedure is provided in Figure 4.2. Participants completed the two-phase study online and were randomly assigned to one of the 8 between-participant groups. Participants were told that the study was about the "perception of events" and asked to comply with several requirements during the experiment (e.g., "please complete the experiment in a single sitting, and do not stop the experiment to complete other tasks"). In Phase 1, participants watched the mock crime video which was followed by two attention check questions and a 3-minute filler task of solving anagrams.

In Phase 2, participants completed the surprise 4AFC memory test. They did not receive any feedback on their memory performance. Participants rated their confidence on a scale from 25% (*not at all confident*) to 100% (*very confident*) that increased in increments of 5%. They were told that 25% was the level of accuracy expected from guessing. We used a 25-100% scale to reflect the minimum to maximum expected frequency of a correct response on a 4AFC memory test (Bornstein & Zickafoose, 1999; Tekin & Roediger, 2017). This enabled us to analyse the correspondence between the expected and observed frequencies of a correct response with calibration statistics. Immediate-confidence participants rated their confidence immediately after answering each question whereas delayed-confidence participants rated their confidence after completing the entire memory test. Delayed-

confidence participants were reminded of their responses as they rated their confidence for each question (e.g., “Question: What was the colour of the stolen bag? Your response: Blue. How confident are you that your response is correct?”). Participants were then asked if they experienced any technical difficulties while watching the video, which device they were using, and if they had complied with the criteria outlined in Phase 1. Finally, participants answered some demographic questions and were debriefed.

We pilot tested the materials to examine whether the night videos produced a lower level of memory accuracy than the day videos, and to check that accuracy exceeded chance performance (i.e., 25%) on each question in the memory tests. The first pilot test was the same as the final experiment, except a brighter version of the night videos was used and the memory tests contained 20 questions. The results of the first pilot experiment (N = 30) revealed that accuracy (i.e., the proportion of correct responses) on the memory test was similar in the night-visibility condition (M = 62.67) and the day-visibility condition (M = 63.00). Therefore, we edited the night videos again and reduced the brightness of the scenarios. In both the car theft and mugging memory tests, 2 out of 20 questions produced below chance accuracy (i.e., 25% correct) and were dropped from all future experiments. To test whether the adjustment to the night videos was sufficient to impair performance in the night-visibility condition, we conducted another pilot experiment. The results of the second pilot experiment (N = 17) revealed that night-visibility participants (M = 62.50) were less accurate on the memory test than day-visibility participants (M = 68.75). Thus, the darker version of the night video was used in the final experiment.

Results

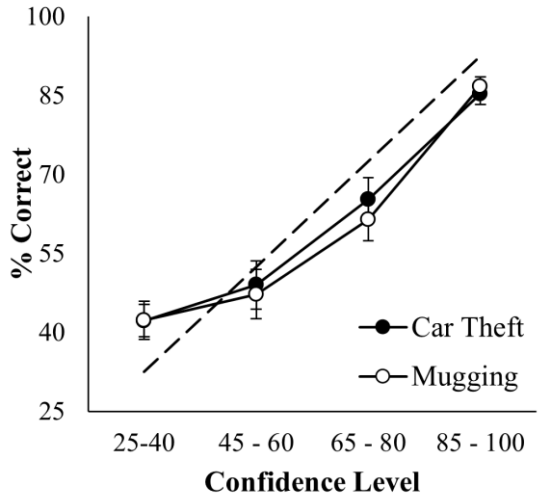
Preliminary Analyses

Before turning to our main analyses, we conducted two preliminary analyses. First, we checked whether calibration was similar across the two events by plotting calibration separately for the mugging and car theft scenarios. Calibration plots were created by plotting accuracy (i.e., the proportion correct) against four levels of confidence (25-40, 45-60, 65-80, and 85-100). The calibration curve in Figure 4.3 shows that calibration did not significantly differ between the two events. Table 4.1 shows the count data for all calibration plots in Experiment 1 and Table 4.2 shows

the calibration statistics. Second, we checked that memory accuracy was lower in the night condition than in the day condition, and the means and non-overlapping CIs in Table 4.3 indicate that it was.

Figure 4.3

Calibration Plot for the Mock Crime Events in Experiment 1



Note. The dashed line represents perfect calibration. The error bars denote the 95% CI around the mean.

Table 4.1

Count Data for Each Calibration Plot in Experiment 1

Condition	Confidence Level			
	25 – 40	45 – 60	65 – 80	85 – 100
Event				
Car Theft	995	459	518	1124
Mugging	700	436	565	1323
Confidence Timing				
Delayed	880	447	562	1243
Immediate	815	448	521	1204
Visibility				
Day	735	419	536	1334
Night	960	476	547	1113

Table 4.2

Calibration Statistics and 95% Confidence Intervals for Experiment 1

Condition	*C*	*OU*
Event		
Car Theft	.064 [.055, .073]	-.233 [-.249, -.216]
Mugging	.050 [.042, .057]	-.205 [-.220, -.189]
Confidence Timing		
Delayed	.060 [.052, .068]	-.224 [-.240, -.209]
Immediate	.054 [.046, .062]	-.213 [-.229, -.197]
Visibility		
Day	.061 [.052, .070]	-.227 [-.242, -.211]
Night	.053 [.045, .061]	-.211 [-.227, -.195]

Note. The *C* statistic reflects the amount of deviation from perfect calibration and ranges from 0 (perfect calibration) to 1. The *OU* statistic reflects the amount of over/underconfidence in participants' responses and ranges from -1 *(complete underconfidence*) to 1 (*complete overconfidence*).

Table 4.3

Means, Standard Deviations and 95% Confidence Intervals for Accuracy and Confidence in Experiment 1

Condition	Accuracy			Confidence		
	M	*SD*	95% CI	*M*	*SD*	95% CI
Event						
Car Theft	62.73	13.60	60.69, 64.76	64.47	13.02	62.53, 66.42
Mugging	66.01	11.22	64.31, 67.70	70.54	11.99	68.72, 72.35
Confidence Timing						
Delayed	64.81	13.04	62.88, 66.75	67.40	12.59	65.53, 69.27
Immediate	63.86	12.09	62.02, 65.69	67.55	13.19	65.54, 69.55
Visibility						
Day	67.86	11.34	66.14, 69.57	70.20	12.36	68.33, 72.07
Night	60.92	12.80	59.00, 62.83	64.80	12.82	62.89, 66.72

Main Analyses

Turning to our main question: Is the confidence-accuracy relationship stronger for immediate-confidence judgements than for delayed-confidence judgements? As there was no difference in calibration across events, our main analyses are collapsed across the car theft and mugging scenarios. Accuracy was calculated as the proportion of correct responses: the number of correct responses divided by the total number of responses.

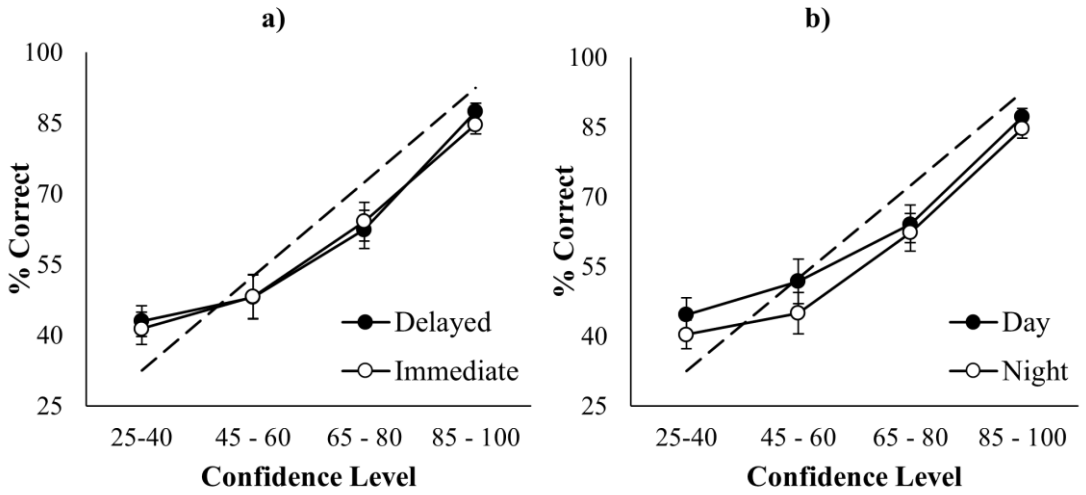
To answer our key question, we conducted a mixed binary logistic regression on response accuracy (i.e., correct vs incorrect) with confidence, confidence timing and visibility as predictors and a random intercept for each participant. This method is preferred to ANOVAs for analysing categorical data, as the latter can yield spurious results (see Jaeger, 2008). Additionally, the inclusion of random intercepts controls for unmeasured variables that are related to memory performance (e.g., fatigue). The model was fitted using the *lme4* package v1.1-23 in R 4.0.2 (Bates et al., 2015) and *p* values were obtained in the *afex* package v. 0.26-1 (Singmann et al., 2018). In line with existing research (e.g., Saraiva et al., 2020; Weber & Brewer,

2008; Wixted & Wells, 2017), the analyses revealed that confidence was a significant predictor of accuracy, $\chi^2(1) = 1003.30$, $p < .001$. The odds ratio was 2.61, indicating that each unit increase in confidence more than doubled the likelihood of a correct answer.

Do immediate-confidence judgements produce a stronger confidence-accuracy relationship than delayed-confidence judgements? Put simply, no. The regression model revealed that there was no significant effect of confidence timing, $\chi^2(1) = 0.94$, $p = .33$, nor was there a significant interaction with confidence, $\chi^2(1) = 0.01$, $p = .92$. To examine the extent to which immediate- and delayed-confidence judgements deviated from perfect calibration, we plotted calibration curves separately for the two confidence timing conditions. Figure 4.4a shows that immediate- and delayed-confidence judgements were similarly calibrated at every level of confidence, such that, (1) accuracy increased as confidence level increased, and (2) there was underconfidence at the lowest level of confidence but overconfidence at higher levels of confidence. Together, these results suggest that participants' confidence judgements provided a reliable indicator of their memory accuracy, regardless of whether these judgements were given immediately after each response or at the end of the memory test.

Figure 4.4

Calibration Plots for a) Confidence Timing and b) Visibility in Experiment 1



Note. The dashed line represents perfect calibration, and the error bars denote the 95% CI around the mean.

Recall that our preliminary analyses showed that memory accuracy was poorer in the night condition than in the day condition, but does visibility influence the strength of the confidence-accuracy relationship? The model confirmed that participants in the day condition were more likely to be accurate than participants in the night condition, $\chi^2(1) = 5.46$, $p = .02$ ($OR = 1.19$). Specifically, day participants were 19% more likely to be correct than night participants, and visibility did not significantly interact with confidence, $\chi^2(1) = 0.01$, $p = .92$. Finally, as Figure 4.4b shows, the visibility (day vs night scene) manipulation did not affect the strength of the confidence-accuracy relationship. Together, these findings suggest that witnesses do have the ability to adjust their confidence ratings to align with their lower accuracy when they recognise that their memory has been compromised by poor viewing conditions.

Conclusion

In sum, Experiment 1 suggests that the confidence-accuracy relationship is not affected by the timing of confidence judgements or the visibility of the witnessed event. Put simply, the accuracy of participants' reports increased as their confidence

in those reports increased, regardless of whether confidence judgements were taken immediately after each response or at the end of the memory test. Furthermore, although night visibility impaired eyewitness accuracy compared to day visibility, the pattern of calibration was similar regardless of visibility. This finding suggests that participants recognised when their memory had been impaired by poor visibility and adjusted their confidence judgements accordingly.

In Experiment 1, the confidence-accuracy relationship was strong when participants were asked relatively innocuous questions, regardless of the timing of the confidence judgements. In Experiment 2, we aimed to replicate this finding in a cued recall test format and to investigate how misleading questions affect the confidence-accuracy relationship.

Decades of research shows witnesses can come to report misinformation they have gleaned from several sources, and sometimes report it with high confidence (e.g., Flowe et al., 2019; Gabbert et al., 2004, 2012; Loftus, 2005; Loftus et al., 1978; Paterson & Kemp, 2006). For example, witnesses who read a misleading narrative about a witnessed event can come to report details that they never experienced. In one study, participants saw four slides depicting different events (Zaragoza & Lane, 1994). Next, they read a narrative or answered questions which contained some incorrect information about the images they had seen. After a 10-minute filler task, participants were given a source memory test, where they were asked to indicate whether they had seen items in the slides, post-event information, both or neither. Participants were more likely to incorrectly report seeing an item in the slides when they had encountered misinformation about the item (misled items), than when they had not encountered misinformation about the item (control items). Importantly, participants were more likely to be highly confident when they misattributed the source of misled items to the original event, than when they misattributed the source of new, control items (Zaragoza & Lane, 1994). These findings are consistent with a growing body of research showing that exposure to misinformation impairs people's ability to discriminate between correct and incorrect details (Bonham & González-Vallejo, 2009; Cann & Katz, 2005; Flowe et al., 2019; Horry et al., 2014).

Guided by previous research and Koriat's (1997) cue-utilisation theory, we predicted that the confidence-accuracy relationship would break down when witnesses were exposed to misinformation about a witnessed event. As outlined in Chapter 2, cue-utilisation theory suggests that people use experience-based cues such as retrieval fluency to judge their confidence judgements and give higher confidence judgements to details that are recalled quickly, than details that are recalled relatively slowly (Koriat, 1997). Furthermore, research suggests that people often do not notice the discrepancy between the misinformation and the original event and, thus, fail to realise that retrieval fluency is a misleading indicator of accuracy (Horry et al., 2014; D. S. Lindsay & Johnson, 1989). Relying on retrieval fluency can serve to inflate a person's confidence because misinformation is encountered more recently than the original event, and as such, should be recalled with more ease. If so, participants should show more overconfidence and poorer calibration when they have been misled about an item than when they have not been misled.

Experiment 2

The aims of Experiment 2 were to replicate the findings on confidence timing in a cued recall task, and to examine how exposure to misinformation affects the confidence-accuracy relationship. To achieve these aims, we used a similar procedure to Experiment 1, except there was a misinformation phase between the event and the final memory test, and the final memory test was a cued recall test. During the misinformation test, each participant answered 4 consistent and 4 misleading questions about the mock crime event they had seen. On consistent questions, participants were provided with the correct response and an incorrect response. On misleading questions, by contrast, they were provided with two incorrect responses. After the misinformation phase, participants completed a memory test that included questions about the critical items they had been asked about during the misinformation phase (i.e., consistent and misled items), and some items that they had not been previously asked about (i.e., control items). As in Experiment 1, participants rated their confidence in each response either immediately after each response (immediate-confidence judgement), or at the end of the memory test (delayed-confidence judgement).

Method

Participants and Design

We used a 2 (Event: car theft, mugging) x 2 (Confidence timing: immediate, delayed) x 3 (Item type: misleading, consistent, control) mixed design with item-type as the within-participants factor. We recruited 346 adults residing in Canada, the United Kingdom and the United States through Amazon's Mechanical Turk using the CloudResearch platform (Litman et al., 2017). Participants received $1.70 for completing the experiment. We excluded participants if they answered an attention check question incorrectly (n = 14), failed to comply with any of the criteria outlined in the experiment (n = 23), experienced technical difficulties (n = 5) or reported suspicions about the misleading questions (n = 24).

Based on previous calibration research, we aimed to collect 12 observations from 280 participants (840 observations per condition) to produce stable calibration curves (Brewer & Wells, 2006). The final sample consisted of 280 participants (153 female, 125 male, 2 undisclosed; M = 41.03, SD = 12.33, range = 19-73). There were 70 participants in each of the 4 between-participant groups, resulting in 3,360 observations overall.

Procedure

Participants completed the three-phase study online and were randomly assigned to one of the 4 between-participant groups. They were told that the study was about "perception of events" and asked to comply with several criteria during the experiment (e.g., "please complete the experiment in a single sitting, and do not stop the experiment to complete other tasks"). In Phase 1 (target event), participants watched either the car theft or mugging video (both with day visibility). We selected 12 critical items from each event. For example, critical items in the car theft video included the name of the parking zone, whereas critical items in the mugging video included the colour of the thief's trousers. The video was followed by two attention check questions (e.g., "Was the victim male or female?") and a 3-minute filler task.

In Phase 2 (misinformation), participants were randomly assigned to answer one of three questionnaires containing 8 x 2AFC questions each referring to a different critical item. They were told that each question had two possible responses,

and that they should select a response for each question. Four questions were misleading and contained two incorrect answers (e.g., "Was the stolen car parked in Parking Zone 4 or Parking Zone D?", correct response = Parking Zone 3). The remaining four questions were not misleading and contained the correct answer and one incorrect answer (e.g., "Was the stolen car parked in Parking Zone 3 or Parking Zone D?"). The critical items were counterbalanced so that – across the three questionnaires – each item appeared once as a misled item and once as a consistent item. That is, a third of participants answered the misleading question about the parking zone, another third answered the consistent question about the parking zone, and the remaining participants did not answer a question about the parking zone. After answering the questionnaire, participants completed a 2.5-minute filler task.

In Phase 3 (memory test), all participants completed the same 12-item cued recall test with one question for each of the 12 critical items (e.g., "According to the sign, which parking zone was the stolen car parked in?", see Appendix B for the full memory test). Eight questions referred to critical items that participants were previously asked about in the misleading questionnaire (4 misled items, 4 consistent items), and 4 questions referred to control items that were not asked about in the misleading questionnaire. Participants were informed that some questions related to details that they were asked about previously and told that they should respond based on their memory of the video, and not on their memory of their previous responses. Participants rated their confidence on an 11-point scale, ranging from 0% (*not at all confident*) to 100% (*very confident*). Immediate-confidence participants rated their confidence immediately after answering each question. Delayed-confidence participants rated their confidence after completing the entire memory test. They were reminded of their responses as they rated their confidence for each question (e.g., "Question: Where did the thief put their phone? Your response: In their pocket. How confident are you in your response?"). Participants were then asked (1) if they experienced any technical difficulties while watching the video, (2) which device they were using, (3) if they had complied with the criteria outlined in Phase 1 and (4) if they had any suspicions about the true purpose of the experiment. Finally, participants answered some demographic questions and were fully debriefed.

An online pilot experiment was conducted to check that the misleading questions reduced participants' accuracy on the final memory test. The results of the

pilot experiment (N = 21) revealed that participants were less accurate on misled items (M = 28.27) than they were on consistent (M = 77.38) or control items (M = 51.49).

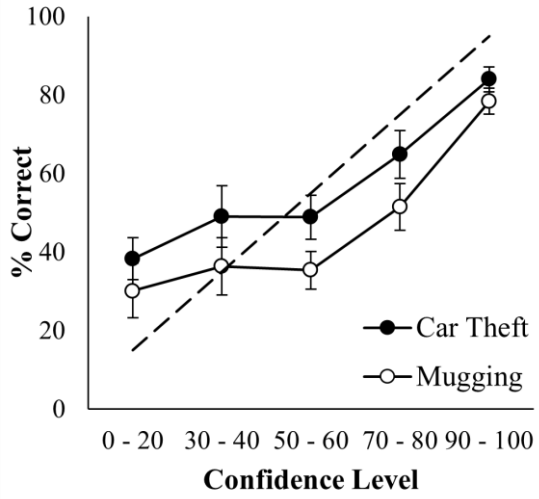
Results

Preliminary Analyses

As in Experiment 1, we checked that calibration was similar for the car theft and mugging scenarios. We created calibration curves by plotting accuracy against 5 levels of confidence (0-20, 30-40, 50-60, 70-80 and 90-100). "I don't know" responses (n = 234, 7% of responses) were excluded and not analysed. Therefore, accuracy was calculated as the number of correct responses divided by the number of correct and incorrect responses (total N = 3,126). The calibration curve in Figure 4.5 reveals that accuracy was lower for the mugging video than for the car theft video, but the overall pattern of calibration was similar for the two events. Table 4.4 shows the count data for all calibration plots in Experiment 2 and Table 4.5 shows the calibration statistics.

Figure 4.5

Calibration Plot for the Mock Crime Events in Experiment 2



Note. The dashed line represents perfect calibration, and the error bars denote the 95% CI around the mean.

Table 4.4

Count Data for Each Calibration Plot in Experiment 2

Condition	Confidence Level				
	0 - 20	30 - 40	50 - 60	70 - 80	90 - 100
Event					
Car Theft	324	159	305	236	518
Mugging	169	168	379	276	592
Confidence Timing					
Delayed	247	153	311	216	604
Immediate	246	174	373	296	506
Item Type					
Control	187	118	172	168	359
Misled	180	112	287	197	280
Consistent	126	97	225	147	471

Table 4.5

Calibration Statistics and 95% Confidence Intervals for Experiment 2

Condition	*C*	*OU*
Confidence Timing		
Delayed	.032 [.024, .040]	.070 [.046, .094]
Immediate	.027 [.020, .034]	.064 [.040, .087]
Event		
Car Theft	.027 [.020, .035]	-.002 [-.026, .022]
Mugging	.036 [.027, .044]	.134 [.110, .157]
Item Type		
Control	.021 [.013, .029]	.036 [.006, .065]
Misled	.087 [.071, .104]	.228 [.198, .258]
Consistent	.034 [.022, .045]	-.063 [-.089, -.037]

Note. The *C* statistic ranges from 0 (*perfect calibration*) to 1 and the *OU* statistic ranges from -1 (*complete underconfidence*) to 1 (*complete overconfidence*).

We also checked that accuracy was lower for misled items than for consistent and control items, and the non-overlapping CIs presented in Table 4.6 show that it was. Accuracy was also significantly higher for consistent items than control items. Notably, although accuracy was lower for misled items than control items, there was no significant difference in confidence between these item types. This suggests that participants failed to adjust their confidence judgements to compensate for their lower memory accuracy after they been exposed to misinformation.

Table 4.6

Means, Standard Deviations and 95% Confidence Intervals for Accuracy and Confidence in Experiment 2

Condition	Accuracy			Confidence		
	M	*SD*	95% CI	*M*	*SD*	95% CI
Event						
Car Theft	60.45	14.58	58.04, 62.87	60.04	17.29	57.17, 62.90
Mugging	54.27	17.04	51.45, 57.09	65.96	16.35	63.26, 68.67
Confidence Timing						
Delayed	57.78	15.58	55.2, 60.36	63.85	17.20	61.00, 66.70
Immediate	56.94	16.70	54.18, 59.71	62.15	16.93	59.34, 64.95
Item Type						
Control	59.23	27.88	55.96, 62.5	62.63	22.92	59.94, 65.32
Misled	36.99	26.67	33.86, 40.13	59.56	21.31	57.05, 62.06
Consistent	75.81	24.51	72.93, 78.68	69.78	19.36	67.51, 72.05

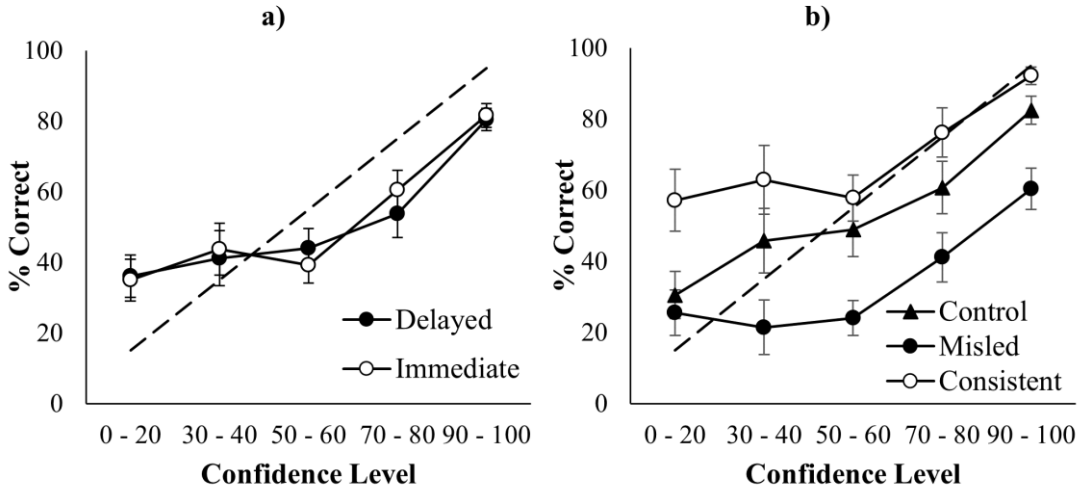
Main Analyses

Recall that the main aims of Experiment 2 were to replicate the findings for confidence timing and to investigate the influence of misinformation on the confidence-accuracy relationship. To address these aims, we collapsed the data across events then conducted a binary logistic regression on response accuracy (correct vs incorrect) as in Experiment 1, with confidence, confidence timing, and item type as predictors. The model confirmed that confidence was a significant

predictor of accuracy, $\chi^2(1) = 392.97$, $p < .001$. The odds ratio was 2.57, very similar to Experiment 1 (2.61), indicating that each point increase in confidence doubled the odds of a correct response. Consistent with Experiment 1, the regression model revealed that there was no significant effect of confidence timing, $\chi^2(1) = 0.01$, $p = .91$, nor did confidence timing significantly interact with confidence, $\chi^2(1) = 0.52$, $p = .47$. Furthermore, the calibration plot in Figure 4.6a shows that immediate- and delayed-confidence judgements produced similar calibration at every level of confidence. Thus, the time at which participants provided confidence judgements did not significantly influence the confidence-accuracy relationship for cued recall.

Figure 4.6

Calibration Plots for a) Confidence Timing and b) Item Type in Experiment 2



Note. The dashed line represents perfect calibration, and the error bars represent the 95% CI around the mean.

Is the confidence-accuracy relationship weaker for misled items than for consistent and control items? The regression model revealed that item type was a significant predictor of accuracy, $\chi^2(2) = 304.03$, $p < .001$. Participants were 64% less likely to be correct for misled items ($OR = 0.36$) and 106% more likely to be correct for consistent items than control items ($OR = 2.06$), and item type did not significantly interact with confidence, $\chi^2(2) = 2.83$, $p = .24$.

To what extent did different item types deviate from perfect calibration? Figure 4.6b shows that misled items produced significant overconfidence at almost every level of confidence. Consistent items, in contrast, produced significant underconfidence at the two lowest levels of confidence, but the best calibration at higher levels of confidence. Consistent questioning may have strengthened participants' memory for the target items, which increased accuracy with a smaller effect on confidence. Finally, control items produced the most linear increase in accuracy across increasing levels of confidence, resulting in underconfidence at the lowest levels of confidence and overconfidence at the highest levels of confidence.

Conclusion

In sum, Experiment 2 supports the finding that the timing of confidence judgements does not affect the confidence-accuracy relationship. Furthermore, consistent with previous research (e.g., Flowe et al., 2019), misinformation impaired the confidence-accuracy relationship, producing overconfidence at almost every level of confidence. This finding supports the idea that people, at least partly, rely on retrieval fluency to judge the accuracy of their memories.

Experiments 1 and 2 suggest that the timing of participants' confidence judgements did not affect the confidence-accuracy relationship in recognition (Experiment 1) or cued recall (Experiment 2) tasks. Recall that participants were asked to answer a relatively small number of questions, such that they are likely to have completed the memory test relatively quickly. Real eyewitnesses, by contrast, are often subjected to lengthy interviews in which they are encouraged to freely report everything that they can remember. It remains possible that our findings might underestimate the impact of delaying confidence judgements in real police interviews, where the delay between giving a response and its corresponding confidence judgement is likely to be relatively long. Therefore, in Experiment 3, we aimed to replicate our findings in a more forensically relevant recall task. We also examined whether the timing of confidence judgements influences the completeness of eyewitnesses' memory reports. Asking participants for confidence judgements immediately after each response could disrupt their narration and the flow of recall (Fisher & Geiselman, 2010). As a result, we might expect immediate-confidence

participants to provide shorter reports (i.e., recall fewer details) than delayed-confidence participants.

Chapter 5:
Re-Examining the Influence of Confidence Timing

Experiment 3

In Experiment 3, we aimed to replicate the finding that confidence timing does not affect the confidence-accuracy relationship in a more forensically relevant recall task. To investigate this, participants watched one mock crime video, completed free recall and cued recall tests for details in the video, and then completed this procedure again for an alternative mock crime video. Research shows that eyewitness free reports are typically associated with high accuracy at every level of confidence, and responses are mostly given with high confidence (Saraiva et al., 2020). Therefore, we included cued recall questions to produce enough variation in accuracy and confidence to plot stable calibration curves.

Method

Participants & Design

We used a 2 (Event: car theft, mugging) x 2 (Confidence timing: immediate, delayed) mixed design with event as the within-participants factor. As people rarely report responses held with low confidence under free report conditions, this study required more observations to achieve stable calibration estimates. Therefore, we aimed to collect 10 observations from at least 200 participants, producing at least 1,000 observations in each between-participants condition. In total, we collected data from 238 participants. Of these participants, 146 were first year Psychology students at the University of Warwick who participated in partial fulfilment of course requirements. The remaining 92 participants were recruited from the wider university community and received £6 upon completing the experiment. We excluded those who did not complete the recall tasks correctly (e.g., if they did not provide confidence ratings, $n = 20$) or reported technical difficulties ($n = 7$). The final sample consisted of 211 participants (43 male, 164 female, 4 undisclosed, M = 19.77 years, SD = 3.05, range = 18-40), including 111 in the immediate-confidence condition and 100 in the delayed-confidence condition.

Procedure

The study was conducted in a computer lab. Participants took part in small groups of 2-10 people, but each participant was seated at their own computer with a set of headphones. Participants were told that the study was about the "perception of events." The entire group was randomly assigned to either the immediate-confidence or delayed-confidence condition. In Phase 1, participants were shown one of the mock crime videos used in Experiment 2 and told to watch it carefully. Next, the 3-minute filler task began.

In Phase 2, participants completed the free recall task. Instructions were presented on the computer screen and participants completed the test in a paper response booklet. Participants were instructed to write down everything that they could remember about the video and to write each detail on a new line. They were told that they could vary the level of detail in their responses and given examples of general-level responses (e.g., "there were 3 to 6 people in the shop") and specific-level responses (e.g., "there were 4 people in the shop"). The examples were unrelated to the content in the video and showed how specificity could be varied for different types of details. Immediate-confidence participants rated their confidence from 0 (*not at all confident*) to 100 (*very confident*) immediately after writing each response. Delayed-confidence participants were asked to write down everything they could remember and to leave the box next to each response blank.

Next, participants completed a 10-item cued recall test. Each question targeted a different critical item (e.g., "What colour was the thief's coat?"). The questions were presented on a computer screen and participants wrote their answers on a paper answer sheet (see Appendix C for the full cued recall tests). Immediate-confidence participants were reminded to write their confidence immediately for each response. Delayed-confidence participants rated their confidence only after completing both the free recall and cued recall tests. They were told to look over their responses and rate their confidence for each response in the free recall task and then the cued recall task.

This two-phase procedure was then repeated for the alternate video. The order of the video was counterbalanced. Finally, participants were asked if they

experienced any technical difficulties while watching either of the videos, answered some demographic questions and were fully debriefed.

Data Coding

Participants' written responses were coded as correct, incorrect, or not applicable ("don't know"). Following Vredeveldt and Sauer (2015), responses were coded as incorrect if they contained any incorrect details even if they were partly accurate. For example, if the thief put money into a *blue* backpack then "the thief put money in a *green* backpack" would be coded as incorrect. Responses were also coded for specificity. Specific answers included a precise description of an item, person or location in the video (e.g., "the thief wore blue skinny jeans"), whereas general answers included only a broad, imprecise description (e.g., "the thief wore jeans"). When there was no clear distinction between general and specific, responses were coded as not applicable. All responses were coded by two independent coders who were blind to participants' conditions, and percentage agreement exceeded 89% for accuracy ($M = 0.89$, $\kappa = .71$, $p < .001$) and specificity ($M = 0.90$, $\kappa = .84$, $p < .001$). The same raters then discussed discrepancies to reach total agreement.

Table 5.1

Count Data for Each Calibration Plot in Experiment 3

Condition	Confidence Level				
	0 - 20	30 - 40	50 - 60	70 - 80	90 - 100
Event					
Car Theft	236	196	378	533	3668
Mugging	176	240	453	754	4570
Video					
A	250	278	500	703	3785
B	162	158	331	584	4453
Confidence Timing					
Delayed	215	215	395	599	4000
Immediate	197	221	436	688	4238
Specificity					
General	119	105	245	441	3619
Specific	225	262	464	727	4061

Note. The specificity calibration plots included a smaller number of observations (n = 10, 268) because some responses could not be coded as general or specific (n = 936) and were removed.

Results

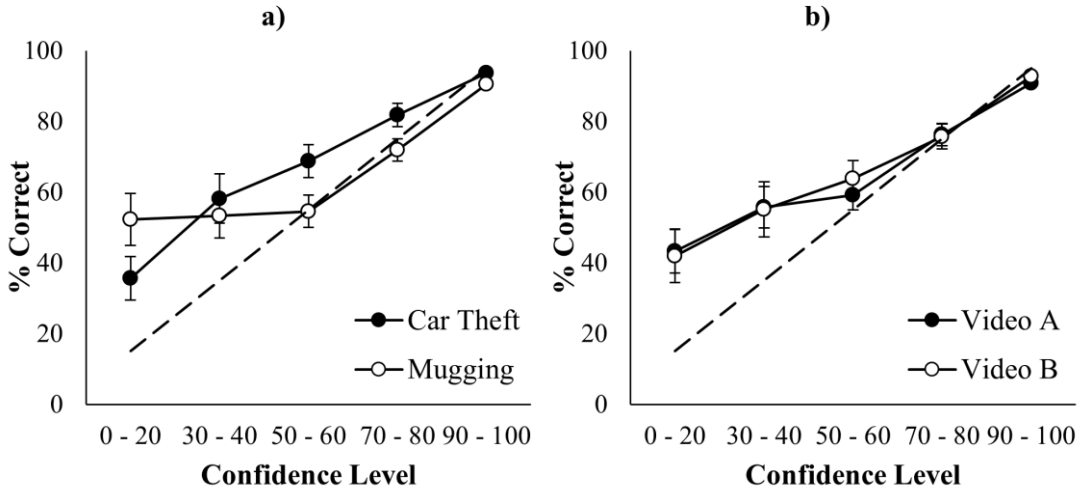
Preliminary Analyses

As in Experiments 1 and 2, we checked that calibration was similar for the car theft and mugging scenarios. Responses coded as NA for accuracy or confidence (i.e., when confidence could not be read) were removed from the analysis (n = 1,160). In total, 11,204 responses were included in our analysis and the count data for each calibration plot in Experiment 3 are presented in Table 5.1. Figure 5.1a shows that accuracy was higher for the car theft video except at the lowest level of confidence, but the pattern of calibration was similar across events (see Table 5.2 for all the calibration statistics for Experiment 3). Next, we checked that calibration was similar for the first and second video, and Figure 5.1b shows that it was. There were

no significant differences at any level of confidence.

Figure 5.1

Calibration Plots for a) the Car Theft and Mugging Events and b) the First (Video A) and Second (Video B) Video in Experiment 3



Note. The dashed line represents perfect calibration, and the error bars represent the 95% CI around the mean.

Table 5.2

Calibration Statistics and 95% Confidence Intervals for Experiment 3

Condition	*C*	*OU*
Event		
Car Theft	.008 [.006, .011]	.001 [-.008, .010]
Mugging	.011 [.009, .013]	.048 [.039, .057]
Video		
A	.011 [.008, .013]	.025 [.015, .035]
B	.007 [.005, .009]	.029 [.021, .038]
Confidence Timing		
Delayed	.011 [.009, .014]	.018 [.009, .028]
Immediate	.007 [.005, .009]	.036 [.027, .044]
Specificity		
General	.012 [.009, .015]	-.009 [-.018, -.001]
Specific	.009 [.008, .011]	.062 [.052, .072]

Note. The *C* statistic reflects the amount of deviation from perfect calibration and ranges from 0 (*perfect calibration*) to 1. The *OU* statistic reflects the amount of over/underconfidence in participants' responses and ranges from -1 (*complete underconfidence*) to 1 (*complete overconfidence*).

Main Analyses

Turning to our main research question: Is the confidence-accuracy relationship affected by the timing of confidence judgements in eyewitness recall? To answer this question, we collapsed the data over event type and video order, then conducted a binary logistic regression on response accuracy (correct vs incorrect) with confidence and confidence timing as predictors. To produce enough variation in accuracy and confidence to plot stable calibration curves, we aggregated each participant's responses in the cued recall and free recall tests. Summary statistics are provided in Table 5.3. As in Experiments 1 and 2, confidence was a significant predictor of accuracy, χ^2 (1) = 1332.30 (*OR* = 2.37), $p < .001$, such that the odds of a correct response doubled with each unit increase in confidence.

Table 5.3

Means, Standard Deviations and 95% Confidence Intervals for Accuracy and Confidence in Experiment 3

Condition	Accuracy			Confidence		
	M	*SD*	95% CI	*M*	*SD*	95% CI
Event						
Car Theft	85.94	9.27	84.69, 87.2	85.66	8.48	84.52, 86.81
Mugging	81.98	9.56	80.69, 83.27	86.91	8.26	85.8, 88.03
Video						
A	81.60	10.26	80.21, 82.98	83.57	8.71	82.4, 84.75
B	86.33	8.29	85.21, 87.44	89.00	7.09	88.04, 89.96
Memory Test						
Cued Recall	74.08	10.82	72.62, 75.54	74.47	10.98	72.99, 75.95
Free Recall	89.66	6.89	88.73, 90.59	93.61	5.88	92.82, 94.41
Confidence Timing						
Delayed	84.89	5.89	83.74, 86.05	86.43	7.19	85.03, 87.84
Immediate	82.92	7.88	81.45, 84.39	86.49	7.42	85.11, 87.87
Specificity						
General	90.59	9.09	89.36, 91.81	89.12	9.04	87.9, 90.34
Specific	78.95	9.53	77.67, 80.24	84.92	8.70	83.75, 86.09

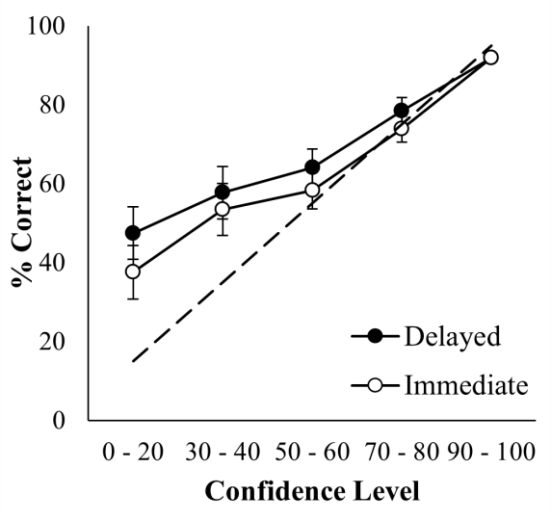
The regression model revealed, once again, that confidence timing did not significantly predict accuracy, $\chi^2(1) = 3.49$, $p = .06$. Due to the limited number of observations at low levels of confidence (0-20), we did not include a confidence x confidence timing interaction in the regression model. However**,** Figure 5.2 shows that calibration was similar between the two groups.

We predicted that delayed-confidence participants might provide more complete reports than immediate-confidence participants. To test this hypothesis, we calculated completeness as the number of free recall responses given by each participant and conducted a one-way ANOVA on completeness using the *aov* package in R. This revealed that the timing of confidence judgements did not significantly influence the completeness of participants' reports, $F(1, 209) = 1.75$, p

= .19. For a summary of the free recall data, see Table 5.4.

Figure 5.2

Calibration Plot for Immediate- and Delayed-Confidence Judgements in Experiment 3



Note. The dashed line represents perfect calibration. The error bars denote the 95% CI around the mean.

Table 5.4

Means and Standard Deviations for Completeness by Confidence Timing in Experiment 3

Condition	Correct Details Reported		Incorrect Details Reported		Total Details Reported	
	M	*SD*	*M*	*SD*	*M*	*SD*
Delayed	32.51	10.92	3.59	2.36	35.96	11.79
Immediate	30.19	12.39	3.75	2.45	33.67	13.21

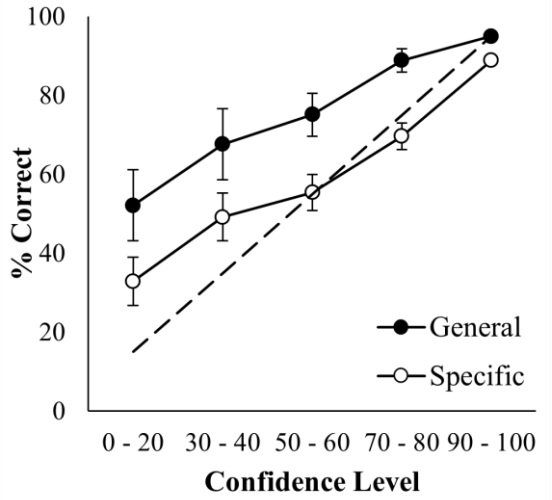
Exploratory Analysis

Previous research shows that specific or "fine-grain" responses (e.g., "the man was wearing a black coat") are typically less accurate but reported with higher confidence than general or "coarse-grain" responses (e.g., "the man was wearing a dark coat"; Goldsmith et al., 2002; Koriat & Goldsmith, 1996). To examine the extent to which specificity predicts the accuracy of eyewitness reports, we conducted a logistic regression on accuracy with confidence and specificity as predictors. Responses coded as NA for specificity were removed (n = 936) and the remaining 10,268 responses were included in the analysis. Each unit increase in confidence doubled the likelihood of a correct response, χ^2 (1) = 1113.38, p < .001 (OR = 2.31), and general responses were 61% more likely to be correct than specific responses, χ^2 (1) = 204.32, p < .001 (OR = 0.39).

Several studies show that specific details are associated with greater overconfidence than general details, but less is known about how specificity affects' eyewitness calibration (Brewer et al., 2018; Vredeveldt & Sauer, 2015). Figure 5.3 shows that calibration was better for specific details than general details. In line with previous research, specific details produced strong calibration and only a small amount of overconfidence at the highest level of confidence. General details, however, produced poorer calibration and significant underconfidence at every level of confidence except the highest level.

Figure 5.3

Calibration Plot for General and Specific Responses in Experiment 3



Note. The dashed line represents perfect calibration, and the error bars denote the 95% CI around the mean.

Discussion

In Experiments 1-3, we investigated the relationship between witness memory accuracy and confidence, and specifically, how the timing of confidence judgements influences this association. We found that the confidence-accuracy relationship was reasonably strong across three different memory test formats (recognition, cued recall, and free recall), and not significantly affected by the timing of confidence judgements. When participants were exposed to misinformation, however, the confidence-accuracy relationship was substantially impaired. While exposure to consistent questions produced a strong confidence-accuracy relationship, exposure to misleading questions produced high levels of overconfidence at almost every level of confidence.

Given there were good reasons to predict that delaying witnesses' confidence judgements should impair the confidence-accuracy relationship, why did we not observe any effect of confidence timing here? One possible explanation, guided by Koriat's (1997) cue-utilisation theory, is that people use a variety of memorial cues to make their confidence responses, including their own meta-memorial beliefs. That

is, people have common-sense—but not necessarily accurate—beliefs about what factors do and do not affect memory accuracy, and people use these beliefs to decide whether to be relatively cautious or not when making confidence judgements. As outlined in Chapter 2, the relationship between confidence and accuracy may be disrupted when witnesses' fail to realise that their memory has been affected by a given factor, and as such, fail to appropriately adjust their confidence ratings. This mechanism could explain why delaying participants' confidence ratings did not affect confidence-accuracy calibration. In Chapter 3, we predicted that delayed-confidence participants might experience increased processing fluency, because they reviewed their memory responses—a form of re-exposure—while providing confidence ratings, and this may serve to inflate their confidence judgements. It is possible, however, that delayed-confidence participants correctly recognised the source of that increased fluency (indeed, we reminded them that they were viewing their own answers once again) and therefore deliberately made the appropriate adjustments in confidence.

The accumulating data on confidence-accuracy calibration in witness interviews appears to fit with this meta-memorial account. Factors that have been shown to disrupt calibration such as exposure to misinformation (Flowe et al., 2019) and marijuana intoxication (Pezdek et al., 2020) may be factors for which people tend to hold misguided meta-memorial beliefs. For instance, a witness might not realize they have been exposed to misinformation or they may underestimate the impact of misinformation on memory and therefore see no reason to adjust their confidence. Conversely, other factors have been shown to have little or no impact on calibration, including weapon focus (Carlson et al., 2017), stress (Pezdek et al., 2021), and attention (Sauer & Hope, 2016). These may be factors for which people tend to hold more accurate meta-memorial beliefs. We should note that we did not set out to test underlying mechanisms here, but a greater theoretical understanding of calibration in witness interviews will be essential to advancing procedures that maximise calibration, and to predicting a priori when the confidence-accuracy relationship is likely to be impaired.

We predicted that asking participants to provide confidence ratings immediately after reporting each detail, rather than at the end of the retrieval process, could lead to fewer details being reported. Surprisingly, we found that the timing of

confidence judgements did not affect the number of details that eyewitnesses reported (Experiment 3). One potential explanation for this is that immediate-confidence participants were able to finish writing each response before providing a confidence judgement, so their reporting was not interrupted or cut short. Even though we did not observe any effect of confidence timing on the completeness of participants' reports here, there may be good reasons not to collect confidence judgements immediately after each reported detail in real world investigative interviews. In real interviews, police often stop witnesses mid-response, which may frustrate witnesses and lead them to give fewer or shorter responses (Fisher & Geiselman, 2010). Immediate-confidence judgements may also provide a challenge for investigative interviewers, as it may be difficult for the interviewer to determine when it is appropriate to pause the witness without reducing the completeness of their report. Deciding exactly when to interrupt a witness may also increase the interviewer's cognitive load, which could in turn impair the interviewer's ability to accurately recall the information being reported by the witness and to ask effective follow-up questions (Hanway et al., 2021).

Consistent with previous research, our data showed that factors that impair memory accuracy do not necessarily reduce the confidence-accuracy relationship (Palmer et al., 2013; Wixted & Wells, 2017). In Experiment 1, night visibility reduced the accuracy of eyewitness reports, but did not compromise the confidence-accuracy relationship. Participants seemingly reduced their confidence to compensate for their lower memory accuracy, so calibration remained strong. This finding is not so surprising given that people are likely to be well aware of the effect of poor visibility on memory accuracy. Consistent with the meta-memorial account given above, when watching the event under low visibility (night) conditions, participants may have found it more difficult to make out precise details, such as the colour of objects, and subsequently interpreted this reduced processing fluency as a sign that their memory was not highly accurate. Additionally, participants may have appropriately reduced their confidence by applying a specific theory about how visibility affects memory performance (Busey et al., 2000; Leippe et al., 2009). While we cannot distinguish the basis for participants' confidence ratings in this study, our findings highlight that poor viewing conditions do not necessarily make

for an unreliable witness as relatively high accuracy can still be observed at high levels of confidence (Wixted et al., 2018).

Whereas participants seemed to adapt their confidence judgements appropriately under different encoding conditions, they failed to do so when they encountered misleading questions. Decades of research shows that people often report information that is inconsistent with what they experienced, but studies have rarely investigated how misinformation affects calibration (Bonham & González-Vallejo, 2009; Gabbert et al., 2004, 2012; Hope et al., 2008; Loftus et al., 1978; Morgan et al., 2013). When witnesses report misinformation with low confidence in the courtroom, this information is likely to be disregarded by triers of fact who use confidence to gauge the accuracy of a witness's testimony. As such, the misinformation is unlikely to result in a miscarriage of justice. Concerningly, the results of Experiment 2 revealed that when participants were exposed to misinformation, they showed overconfidence in their memory accuracy at almost every level of confidence. These findings highlight concerns about the reliability of eyewitness memory and using confidence ratings in forensic contexts where eyewitnesses may unknowingly encounter misinformation (Wade et al., 2018). Given that it may be impossible to determine when witnesses have been exposed to misinformation, the results of Experiment 2 suggest that interviewers should exercise caution when using confidence ratings for assessing the accuracy of eyewitness reports. They also raise further questions about how the confidence-accuracy relationship is affected by misinformation. For instance, how does misinformation gleaned from a co-witness versus an investigative officer (who is deemed to be authoritative) affect the confidence with which people report misinformation? And can calibration be preserved by warning witnesses that their memories may contain information from numerous sources?

Understanding the mechanisms behind witnesses' overconfidence may help researchers to better predict when the confidence-accuracy relationship will break down. One possibility is that misleading questions produced overconfidence because participants did not scrutinise the source of their memories and instead relied on retrieval fluency—that is, the speed with which information came to mind—to make their confidence judgement (Horry et al., 2014). As noted in Chapter 2, previous work suggests that retrieval fluency provides the rememberer with a good predictor

of accuracy under most conditions, but it is deceptive when people are exposed to misinformation (Koriat, 1997). Misinformation is usually encountered after the to-be-remembered stimuli, so it tends to come to mind more quickly than the originally encoded information giving rise to the mistaken impression that the memory is accurate.

Finally, we found evidence to suggest that repeated questioning can improve memory accuracy in specific contexts (Odinot et al., 2009). Specifically, participants in Experiment 2 showed higher levels of memory accuracy on consistent items than on misled and control items. This finding fits with a growing body of research which suggests that repeated questioning does not produce confidence inflation, as originally thought (Odinot & Wolters, 2006; Odinot et al., 2013; Shaw & McClure, 1996). The novel contribution of Experiment 2, however, is that it shows how repeated questioning influences eyewitness calibration. Specifically, we found that participants were more likely to show underconfidence on consistent items than on control items. Furthermore, participants maintained relatively high accuracy and strong calibration at high levels of confidence regardless of whether they were asked about an item once or twice.

To conclude, Experiments 1-3 help to refine the parameters in which witness confidence serves as a useful indicator of memory accuracy. Whether confidence is collected during or immediately after memory retrieval may have little or no bearing on the confidence-accuracy relationship, but exposure to misinformation can have substantial, detrimental effects to the value of witnesses' confidence ratings. Given that legal decision-makers rely heavily on witnesses' confidence in the courtroom (Brewer & Burke, 2002; Garrett et al., 2020), it is crucial that investigators, judges, and juries are advised of the situations in which confidence can be a problem.

Chapter 6:
Do Long Retention Intervals Impair the Confidence-Accuracy Relationship?

Introduction

As outlined in Chapter 1, decades of research have demonstrated the powerful influence that highly confident witnesses can exert on legal decision makers (Cutler et al., 1988; Garrett et al., 2020; Key et al., 2022). This finding, paired with the need to find reliable ways to assess the reliability of witness evidence, has motivated memory researchers to better understand the relationship between a witness's confidence in their memory and the accuracy of their memory (Palmer et al., 2013; Sporer et al., 1995; Wixted et al., 2018). New research on the witness confidence-accuracy relationship indicates that eyewitness confidence—at least in some contexts—is a reasonably reliable predictor of memory accuracy (Brewer & Wells, 2006; Wixted & Wells, 2017; see also Berkowitz & Frenda, 2018; Wade et al., 2018). For example, one study suggested that lineup identifications made with high confidence are likely to be, on average, highly accurate, even when eyewitness accuracy is reduced by variables outside the control of the legal system (e.g., long retention intervals; Semmler et al., 2018). But there is also evidence that some factors, such as feedback about one's memory ability and exposure to misinformation, may impair the confidence-accuracy relationship (Flowe et al., 2019; Iida et al., 2020; Pezdek et al., 2020). One well-studied factor that is known to systematically reduce the quantity and accuracy of witnesses' memory reports but has rarely been studied in the confidence-accuracy literature, is the delay period between when a crime occurs and when a witness is asked to provide evidence (Tuckey & Brewer, 2003). The aim of Experiment 4 was to examine how different delays affect the confidence-accuracy relationship in eyewitness recall, and to gather information about possible mechanisms of influence.

The few studies that have investigated the influence of retention interval on the confidence-accuracy relationship suggest that longer retention intervals may reduce the correlation between a witness's confidence and their memory accuracy compared to short delays (Horry et al., 2014; Odinot & Wolters, 2006; Odinot et al., 2013). In one study, participants viewed a videotape depicting a car accident, and

following a delay, answered open-ended questions about it. A 5-week delay led to lower confidence in correct responses and a weaker confidence-accuracy correlation ($\gamma = .49$) than a 1-week ($\gamma = .63$) or 3-week delay ($\gamma = .58$, Odinot & Wolters, 2006). Similarly, when participants answered questions about a mock crime after a 1-week delay, they were less able to discriminate between correct and incorrect responses than those who answered questions immediately (Horry et al., 2014). Although these studies provided valuable insight into the impact of different delays on witnesses' confidence judgements, confidence-accuracy calibration – where accuracy is calculated for each level of confidence – was not examined (the sample sizes in Odinot & Walters, $N = 67$, and in Horry et al., $N = 80$, preclude calibration analyses). Thus, it remains unclear whether longer delay periods impair the confidence-accuracy relationship by inducing participants to become more underconfident or overconfident in their memories (Juslin et al., 1996).

To guide our understanding of why and how longer retention intervals might impair confidence-accuracy calibration, we can draw on Koriat's (1997) cue-utilisation theory (outlined in Chapter 2). To recap, cue-utilisation theory suggests that people use multiple cues to make confidence judgements, including information derived from the process of remembering (i.e., experience-based cues) and their personal meta-memorial beliefs (i.e., theory-based cues). One example of an experienced-based cue is the ease with which a target event is recalled. Research shows people tend to assign higher confidence ratings to information that is recalled with relative ease compared to information that requires more cognitive effort to bring to mind (Lindholm et al., 2018). Such experienced-based cues are typically useful indicators of memory accuracy, but sometimes these cues can lead to errors (Horry et al., 2014; Shaw & McClure, 1996). Theory-based cues are, by contrast, any cue related to a person's beliefs about how their memory works. For instance, factors relating to the witnessing situation, such as recalling that the crime was observed in broad daylight, or factors relating to the testing situation such as biased feedback from an interviewer (e.g., "you're spot on"), can lead witnesses to provide more, or less, conservative confidence judgements (Semmler et al., 2004). The usefulness of theory-based cues hinges on whether a witnesses' meta-memorial beliefs are accurate: Recent research suggests that participants can adjust their confidence appropriately for some common-sense factors, such as the influence of

lighting on memory (Experiment 1), but not for other, poorly understood factors, such as the influence of marijuana intoxication on memory (Cormia et al., 2020; Desmarais & Read, 2011; Ost et al., 2017; Pezdek et al., 2020).

Based on Koriat's (1997) cue-utilisation theory, we might expect the length of the retention interval to affect confidence judgements via two mechanisms. First, when witnesses recall details of a target event after a long retention interval, those details are likely to come to mind relatively slowly and they may require a large amount of cognitive effort to recall. Put another way, longer retention intervals should reduce retrieval fluency compared to shorter retention intervals, and thus lead witnesses to lower their confidence judgements (Horry et al., 2014). It is worth noting, though, that while correct details typically come to mind more quickly and fluently than incorrect details, they may come to mind relatively slowly after a long delay (Robinson et al., 1997). As a result, witnesses may struggle to determine the accuracy of their responses following a long delay. Based on this mechanism, we might expect that participants who complete a memory test following a relatively long delay will show poorer confidence-accuracy calibration than those who complete a memory test following a short delay.

Second, people may use their meta-memorial beliefs about how retention intervals affect memory accuracy to guide their confidence judgements (Koriat, 1997; Leippe et al., 2009). Recent research shows that laypeople are generally well aware of the negative effects of long delay periods on memory accuracy (Cormia et al., 2020). As such, witnesses may lower their confidence judgements (appropriately) when their memory is tested following a relatively long delay but not when it is tested after a short delay. If witnesses sufficiently adjust their confidence to reflect their memory accuracy then, based on this mechanism, we might expect delay to have little or no effect on eyewitness confidence-accuracy calibration.

There is, however, at least one factor that could mediate the impact of retention intervals on the confidence-accuracy relationship: people's beliefs about the reliability of their own memory (Leippe et al., 2009). To our knowledge, only one study has examined the influence of self-rated memory ability on witnesses' confidence in their memory reports. In this study, mock witnesses answered several metamemory questionnaires about their general memory ability and then watched a

mock crime video depicting a theft (Saraiva et al., 2020). Five minutes later, participants were asked to recall all the details that they could remember about the video. Self-rated memory ability did not significantly predict the accuracy of witnesses' memory reports or the confidence that they expressed in those reports. Put simply, how people felt about their memory ability did not affect their confidence in their memory performance when they were tested after a relatively short delay.

There are good reasons, however, to expect self-rated memory ability will have a stronger influence on people's confidence judgements when the delay period is long, compared to when the delay period is relatively short. People who believe that they have a highly reliable memory may feel more confident in their ability to accurately remember details following a long delay period than people who believe that they have an unreliable memory. After a short delay, however, even people with low self-rated memory ability may feel relatively confident in their ability to remember details accurately. Thus, we might expect that participants with high self-rated memory ability will be more confident in their memory performance than participants with low self-rated memory ability when the delay is long, but not when the delay is relatively short. Previous research supports this prediction, showing that how people feel about their memory ability affects their confidence in their memory, and this effect is larger when their memory is relatively weak. For example, receiving feedback about one's memory ability has a larger influence on confidence when internal cues are weak than when internal cues are strong (Charman et al., 2010; Iida et al., 2020).

Understanding how retention interval and perceived memory ability affect eyewitness confidence is of practical and theoretical importance. On a practical level, eyewitness confidence is compelling in the courtroom and heavily influences how triers of fact perceive the credibility of eyewitnesses, so it is important to understand when confidence is not a reliable indicator of accuracy (Bradfield & Wells, 2000; Cutler et al., 1988; Fox & Walters, 1986; Grabman et al., 2021; Key et al., 2022; Slane & Dodson, 2022). Additionally, if the confidence-accuracy relationship breaks down over longer retention intervals, then easy-to-administer interviewing tools such as the Self-Administered Interview (SAI; Gabbert et al., 2009) are likely to be important for maintaining the confidence-accuracy relationship in real cases where it is often not possible to interview eyewitnesses immediately. On

a theoretical level, by examining how retention interval and perceived memory ability affect the confidence-accuracy relationship, we may learn more about the cues witnesses tend to rely on when making their confidence judgements.

In Experiment 4, we examined how retention interval affects the confidence-accuracy relationship for eyewitness reports, and whether this depends on how people feel about their memory ability. In Phase 1, participants completed questionnaires about their general memory ability and then watched a mock crime video. Some participants proceeded to Phase 2 immediately, whereas other participants returned to complete Phase 2 after 1 week or 1 month. In Phase 2, participants answered cued recall questions about the video they saw in Phase 1 and rated their confidence in their responses. We predicted that people's beliefs about their memory ability would have a larger impact on confidence when the retention interval was long than when the retention interval was relatively short.

Experiment 4

Method

Participants & Design

We used a 2 (Event: car theft, mugging) x 3 (Delay: immediate, short, long) between-participants design. There are currently no clear guidelines on sample size estimation for individual differences research, so we based our sample size on previous studies which used the same metamemory and personality scales (Jackson & Kleitman, 2014; Kleitman et al., 2019; Saraiva et al., 2020; Zhu et al., 2010). We aimed to recruit at least 600 people (200 per delay condition), producing ~3,000 observations in each delay condition.

In total, we recruited 778 adults through Amazon's Mechanical Turk (MTurk) using the CloudResearch platform (Litman et al., 2017). Participants received $2.50 upon completing the experiment. We excluded those who failed an attention check question (n = 37), reported skipping or watching the video more than once (n = 57), reported being distracted during the experiment (n = 36), experienced technical difficulties (n = 6), or did not complete Part 2 within 3 days of receiving the invitation link (n = 25). A further 6 people (3 short delay, 3 long delay) were excluded because they failed to answer any of the memory test questions correctly.

The final sample consisted of 611 adults (319 women, 286 men, 6 undisclosed, *M* = 42.79 years, *SD* = 12.69, range = 18-83). There were 202 people in each of the immediate and short conditions, and 207 in the long condition, producing 9,165 observations in total. The Department of Psychology Research Ethics Committee at the University of Warwick approved this research. This study was pre-registered; the numeric data and corresponding R code are available on the Open Science Framework: https://osf.io/sgb4z/.

Materials

Multifactorial Memory Questionnaire (MMQ; Troyer & Rich, 2002). The MMQ consists of three scales that measure different aspects of metamemory: Satisfaction (i.e., contentment with one's memory ability; α = 0.91), Strategy (i.e., use of memory strategies in everyday life; α = 0.84) and Ability (i.e., perception of one's memory ability; α = 0.89). MMQ-Satisfaction consists of 18 statements (e.g., "I am generally pleased with my memory ability") rated on a 5-point scale ranging from "*Strongly agree"* to "*Strongly Disagree"*, with higher scores indicating greater contentment with one's memory ability. MMQ-Strategy consists of 19 statements referring to different memory strategies (e.g., "Use a timer or alarm to remind you to do something") and respondents rate how often they use each memory strategy on a 5-point scale ranging from "*All the time"* to "*Never"*, with higher scores indicating more frequent use of memory strategies. MMQ-Ability consists of 20 statements and respondents rate how often they have experienced different memory mistakes (e.g., "Forget to pay a bill on time") on a 5-point scale ranging from "*All the time*" to "*Never*", with higher scores indicating better self-rated memory ability.

Narcissistic Personality Inventory (NPI-16; Ames et al., 2006). The NPI-16 consists of 16 pairs of statements, each including a narcissistic statement (e.g., "I like to be the centre of attention") and a non-narcissistic statement (e.g., "It makes me uncomfortable to be the centre of attention"; α = .72). Respondents indicate the statement that describes them best, with higher scores indicating higher levels of narcissism. We included the NPI-16 because we were initially interested in whether individual differences in personality would affect the amount of under- or overconfidence in witness's memory reports. As the analyses including the NPI-16

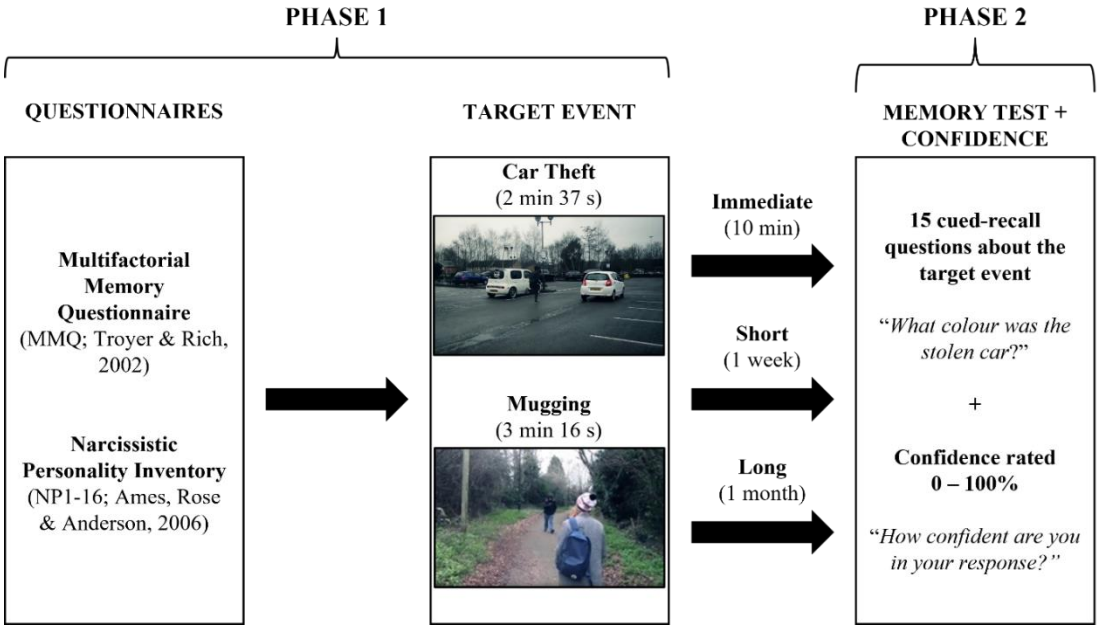
are not relevant to our main research question, we report these results in Appendix D.

Videos. When trying to detect reliable and generalizable effects in witness memory research, it is important to create some variability in the encoding conditions. To this end, we used the same mock crime videos as in Experiments 1-3: a car theft and a mugging scenario. In the car theft scenario (2 min 37 s), a woman scopes out a supermarket car park and notices the victim leaving their car. The victim sees the thief stealing the car and chases after it. In the mugging scenario (3 min 16 s), a woman meets a man and they exchange phone numbers. When the man leaves for class, a thief approaches the woman and wrestles her bag from her before fleeing.

Memory Test. We created a memory test for each video that contained 15 cued recall questions (see Appendix E for the full memory test). Questions varied in difficulty and asked about people, objects, and locations in the video (e.g., "What colour was the stolen car?"). Participants responded by typing their answers into a text box.

Figure 6.1

The General Procedure for Experiment 4



Procedure

An overview of the general procedure is shown in Figure 6.1. Participants completed the study online and were randomly assigned to one of the 6 between-participant conditions. Participants were not told which condition they had been assigned to. They were told that the study was exploring people's "cognitive ability and beliefs about cognition," and were asked to comply with several requirements during the experiment (e.g., "Please do not speak to anyone during the experiment"). In Phase 1, participants completed the metamemory and narcissism measures, and were then randomly assigned to watch one of the two mock crime videos (car theft or mugging). Participants were asked to watch the video carefully, as they would be asked questions about it later. They then answered two attention check questions (e.g., "Which event did you just see?") and were asked if they experienced any technical difficulties while watching the video, if they had watched the entire video, and if they had watched the video only once. Participants then completed a 10-minute filler task of logic problems. Immediate-delay participants proceeded to Phase 2 straight away, whereas short- and long-delay participants were told that they would receive a link either 1 week or 1 month later, respectively.

In Phase 2, participants completed the cued recall memory test and, on the following page, rated their confidence in their responses on a 0-100% scale that increased in increments of 10%. Response time was measured from the time that the question was shown until participants submitted their answer. Participants were not given any feedback on their performance. After completing the memory test, participants were asked if they complied with the criteria set out at the beginning of the experiment and were fully debriefed.

We carried out a pilot experiment (N = 69) to check that participants could still remember details from the mock crime video after a 1-week retention interval. The pilot experiment was the same as the final experiment, except that it did not include a long-delay condition. The results of the pilot test revealed that participants were more accurate when they completed the cued recall memory test immediately (M = 64.23) than when they completed the memory test after 1 week (M = 48.68). Given that both immediate-delay and short-delay participants achieved a reasonable

level of accuracy, we decided to include a long (i.e., 1 month) delay condition in the final experiment.

Data Coding

Participants' answers on the memory test were coded as either correct, incorrect or "don't know". Responses were coded as correct if they correctly described the critical item, regardless of specificity. For example, if the correct answer was "dark blue" then "dark" would also be coded as correct. Responses were coded as "don't know" if participants said that they did not remember the answer, or if they said that they did not notice the detail during the video. Blank responses were not permitted.

Results

Preliminary Analyses

Before conducting the main analyses, we conducted three preliminary analyses. First, we compared the number of "don't know" responses across delay conditions. The number of "don't know" responses increased with delay (immediate-delay n = 163, 5.38%; short-delay n = 181, 5.97%; long-delay n = 254, 8.18%) which suggests that participants may have chosen to withhold more information after longer delays. Specifically, delay was significantly associated with "don't know" responses such that long-delay participants gave more "don't know" responses than immediate-delay participants, $\chi^2(1) = 18.55$, $p < .05$ ($OR = 1.22$), and short-delay participants, $\chi^2(1) = 11.00$, $p < .05$ ($OR = 1.17$). The number of "don't know" responses did not vary significantly between the immediate-delay and short-delay conditions, $\chi^2(1) = 0.89$, $p = .35$ ($OR = 1.05$). This finding squares with existing research that shows witnesses can, and often do, withhold information in an effort to maximise accuracy (Evans & Fisher, 2011; Goldsmith et al., 2002; Weber & Brewer, 2008).

Second, we examined the means for accuracy and confidence across delay conditions. "Don't know" responses were excluded (n = 598, 6.5% of responses) and accuracy was calculated as the number of correct responses divided by the number of correct and incorrect responses (total N = 8,567). The means and non-overlapping 95% confidence intervals presented in Table 6.1 show that short-delay participants

were significantly less accurate than immediate-delay participants and significantly more accurate than long-delay participants. Thus, even though short- and long-delay participants provided more "don't know" responses, the accuracy of their reports was impaired compared to immediate-delay participants. The same pattern was observed for confidence. Short-delay participants were significantly less confident than immediate-delay participants, but significantly more confident than long-delay participants.

Table 6.1

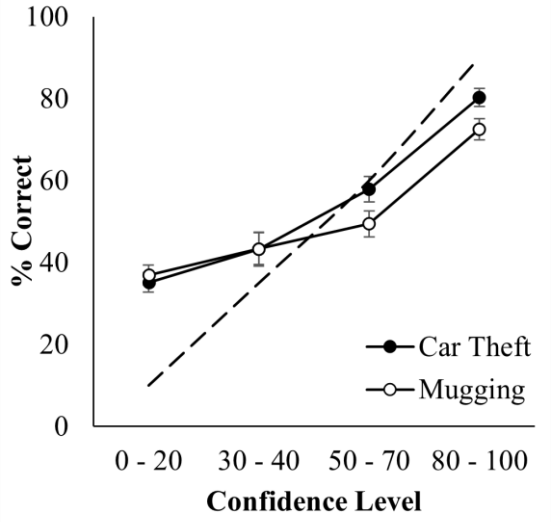
Means, Standard Deviations and 95% Confidence Intervals for Accuracy and Confidence in Experiment 4

Condition	Accuracy			Confidence		
	M	*SD*	95% CI	*M*	*SD*	95% CI
Event						
Car Theft	54.33	18.59	52.24, 56.41	47.50	25.32	44.66, 50.34
Mugging	49.99	18.22	47.95, 52.03	46.20	23.41	43.57, 48.82
Delay						
Immediate	65.83	16.63	63.53, 68.12	69.21	15.30	67.10, 71.32
Short	48.00	15.23	45.90, 50.10	38.83	19.71	36.11, 41.54
Long	42.87	15.25	40.79, 44.95	32.85	20.02	30.12, 35.58

Finally, we checked that calibration was similar for the car theft and mugging events. We created calibration curves by plotting accuracy against 4 levels of confidence (0-20, 30-40, 50-70, 80-100). Figure 6.2 shows that accuracy at 50-70% and 80-100% confidence was higher for the car theft scenario than the mugging scenario, but the overall pattern of calibration was similar for the two events. Table 6.2 shows the count data for all calibration plots in Experiment 4 and Table 6.3 shows the calibration statistics.

Figure 6.2

Results of the Preliminary Analyses for Experiment 4



Note. The dashed line represents perfect calibration, and the error bars denote the 95% CI around the mean.

Table 6.2

Count Data for Each Calibration Plot in Experiment 4

Condition	Confidence Level			
	0 - 20	30 - 40	50 - 70	80 - 100
Event				
Car Theft	1544	552	972	1213
Mugging	1548	624	985	1129
Delay				
Immediate	374	275	684	1534
Short	1257	446	655	491
Long	1461	455	618	317

Table 6.3

Calibration Statistics and 95% Confidence Intervals in Experiment 4

Condition	*C*	*OU*
Event		
Car Theft	.030 [.025, .035]	-.067 [-.082, -.053]
Mugging	.041 [.035, .047]	-.038 [-.054, -.023]
Delay		
Immediate	.024 [.019, .029]	.035 [.018, .052]
Short	.034 [.028, .041]	-.092 [-.111, -.074]
Long	.049 [.041, .057]	-.101 [-.120, -.082]

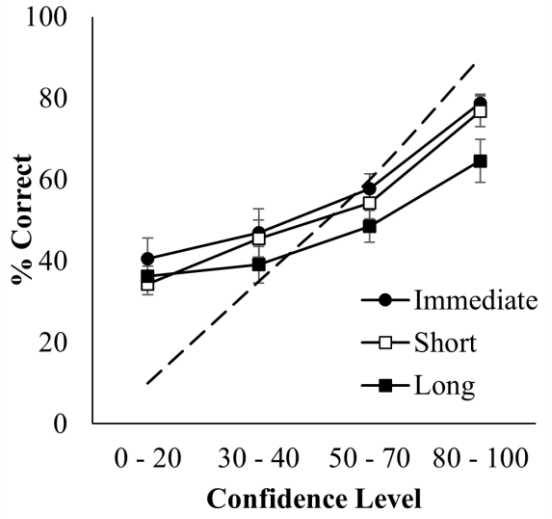
Note. The *C* statistic reflects the amount of deviation from perfect calibration and ranges from 0 (*perfect calibration*) to 1. The *OU* statistic reflects the amount of over/underconfidence in participants' responses and ranges from -1 (*complete underconfidence*) to 1 (*complete overconfidence*).

Main Analyses

Turning to our key research question: How does retention interval influence the confidence-accuracy relationship for eyewitness recall? To answer this question, we collapsed the data over the two events and then plotted calibration curves for each delay condition. The calibration curves were created by plotting accuracy (i.e., the proportion correct) against 4 levels of confidence (0-20, 30-40, 50-70, 80-100) so that each bin contained at least 200 observations. Figure 6.3 shows that the pattern of calibration was generally similar across delay conditions, with some underconfidence at the low levels of confidence and overconfidence at the highest levels of confidence. The only significant difference was at the highest level of confidence. Even though long-delay participants gave relatively few responses with 80-100% confidence (Table 6.2), they were more overconfident in these responses than were short-delay and immediate-delay participants.

Figure 6.3

Calibration for Each Delay Condition in Experiment 4



Note. The dashed line represents perfect calibration. Error bars denote the 95% CI around the mean.

Do people's beliefs about the reliability of their own memory influence their confidence judgements, and if so, does this depend on the length of the retention interval? To answer these questions, we conducted two multiple regressions on mean confidence and over/underconfidence including Delay, MMQ-Satisfaction, MMQ-Ability and MMQ-Strategy as predictors (Table 6.4). We ran the regression models in R 4.0.5 and controlled the false discovery rate by applying the Benjamini-Hochberg correction to all *p* values (Benjamini & Hochberg, 1995). The models were significant for confidence, R^2 =.40, adjusted $p < .001$, and over/underconfidence, R^2 =.09, adjusted $p < .001$. Both short and long delays were associated with lower confidence judgements and greater underconfidence compared to when the memory test was taken immediately. Importantly, this underconfidence was driven by the relatively high number of responses at low levels of confidence in the two delayed conditions.

Table 6.4

Results of the Multiple Regression Models with Delay in Experiment 4

Predictor	Confidence			Over/underconfidence		
	Estimate	SE	Adjusted *p*	Estimate	SE	Adjusted *p*
Long Delay	-34.63	1.90	< .001	-0.13	0.02	< .001
Short Delay	-29.52	1.91	< .001	-0.12	0.02	< .001
MMQ-Satisfaction	2.08	1.96	.72	0.00	0.02	.99
MMQ-Ability	-0.41	1.87	.83	-0.03	0.02	.50
MMQ-Strategy	0.79	1.34	.83	0.01	0.01	.87
Long Delay x MMQ-Satisfaction	-0.85	2.68	.83	0.03	0.03	.50
Short Delay x MMQ-Satisfaction	-2.55	2.70	.72	-0.02	0.03	.84
Long Delay x MMQ-Ability	-0.57	2.70	.83	0.00	0.03	.99
Short Delay x MMQ-Ability	1.20	2.69	.83	0.04	0.03	.42
Long Delay x MMQ-Strategy	-1.81	1.97	.72	0.01	0.02	.87
Short Delay x MMQ-Strategy	1.17	1.93	.83	0.02	0.02	.68

Note. *p* Values were adjusted with the Benjamini-Hochberg procedure.

We suggested that the relationship between retrieval fluency and memory accuracy may break down over time which should lead long-delay participants to show a weaker confidence-accuracy relationship (if their confidence decisions are based, to some extent, on retrieval fluency). Consistent with this mechanism, long-delay participants showed overconfidence at high levels of confidence, but they also gave fewer high confidence responses than did immediate-delay participants, which meant that their overall (mean) decrease in confidence was larger than was justified by their lower (mean) accuracy. None of the metamemory scales predicted

confidence or over/underconfidence, nor did they significantly interact with delay (adjusted $p > .05$). Additional analyses, including narcissism as a predictor of confidence, are reported in Appendix D.

Exploratory Analyses

To further investigate whether the retrieval fluency mechanism could explain the impairment in the confidence-accuracy relationship after a long delay, we used response time as a proxy for retrieval fluency and compared memory accuracy and confidence at different response times for each delay condition. The means presented in Table 6.5 show that response times were generally similar across conditions. To control for between-participant variation, we partitioned response times for each participant into five quantiles and then calculated the mean accuracy and mean confidence for each quantile. Consistent with the notion that people use retrieval fluency to guide their confidence judgements, participants' confidence judgements increased as their response time decreased. That is, confidence was negatively correlated with response quantile in all delay conditions: $\tau = -.23$, $p = < .001$, for the immediate-delay condition, $\tau = -.13$, $p < .001$, for the short-delay condition, and $\tau = -.10$, $p < .001$, for the long-delay condition following Benjamini-Hochberg correction. And consistent with the notion that fluency is associated with accuracy at short, but not longer, delay periods, we found that mean memory accuracy was significantly correlated with response time quantile in the immediate-delay condition, $\tau = -.14$, $p < .001$, but not in the short-delay condition, $\tau = -.04$, $p = .80$, or in the long-delay condition, $\tau = .02$, $p = 1$. Together, these results suggest that people may use retrieval fluency to guide their confidence judgements, even after a delay when it may not be informative about the accuracy of their reports.

Table 6.5

Means, Standard Deviations and 95% Confidence Intervals for Response Time in Experiment 4

Condition	*M*	*SD*	95% CI
Event			
Car Theft	14.39	11.60	13.09, 15.69
Mugging	16.77	21.49	14.36, 19.18
Delay			
Immediate	15.77	11.15	14.23, 17.31
Short	15.47	14.35	13.49, 17.44
Long	15.51	23.75	12.27, 18.75

Discussion

In Experiment 4, we investigated how retention interval affects the relationship between eyewitness memory accuracy and confidence. We found that a 1-month delay impaired the confidence-accuracy relationship compared to a short delay of just 10 minutes. Participants' confidence was significantly associated with retrieval fluency, and not significantly affected by how people felt about their general memory ability, regardless of delay. These findings suggest that although people adjust their confidence judgements to compensate for their reduced memory accuracy after a long retention interval, this adjustment does not necessarily result in a strong confidence-accuracy relationship.

Consistent with previous research (e.g., Odinot & Wolters, 2006; Odinot et al., 2013), the results of Experiment 4 suggest that the accuracy of witnesses' reports decreases as retention interval increases. This finding highlights the importance of collecting eyewitness reports quickly to maximise the accuracy of those reports. Given that it is often not possible for police to interview witnesses immediately after the crime, our results suggest that pen and paper interview tools (e.g., the Self-Administered Interview; Gabbert et al., 2009) may be useful for maximising the accuracy of eyewitness reports. These tools allow investigators to collect a detailed and accurate report from eyewitnesses by administering a booklet which contains

instructions designed to facilitate recall. Research shows that these tools are effective at maintaining eyewitness accuracy over a delay: witnesses who complete the SAI shortly after the crime provide more correct details in their later memory reports than witnesses who do not have an early recall opportunity (Gabbert et al., 2009; Horry et al., 2021). Future research could investigate whether the SAI can also help to enhance the confidence-accuracy relationship after a delay.

When participants were tested after a relatively long delay, they were not only less accurate in their reports but also less able to monitor the accuracy of those reports. Specifically, short- and long-delay participants gave fewer high confidence judgements than immediate-delay participants and expressed underconfidence in their overall memory performance. This is consistent with previous research, showing that the confidence-accuracy relationship is impaired when witnesses are questioned after a relatively long delay (Horry et al., 2014; Odinot & Wolters, 2006; Odinot et al., 2013). Moreover, these findings are consistent with cue-utilisation theory, which suggests that people will reduce their confidence judgements when they believe that their memory performance has been compromised (Koriat, 1997; Leippe et al., 2009).

Although long-delay participants gave fewer high confidence responses (80-100%), they tended to be less accurate and therefore more overconfident in those responses than immediate-delay and short-delay participants. Whereas immediate- and short-delay participants were 78.94% and 76.78% accurate at the highest level of confidence, respectively, long-delay participants were only 64.67% accurate. This finding suggests that high confidence judgements are less likely to be accompanied by high accuracy when memory reports are taken after a relatively long delay. This is important because legal decision makers tend to believe witnesses when they express information with high confidence (Brewer & Burke, 2002; Cutler et al., 1988; Key et al., 2022). Although high confidence errors may be relatively rare when participants are tested after a relatively long delay, the crucial point is that confidence judgements are more informative about the accuracy of eyewitness reports when they are collected shortly after the crime than when they are collected after a relatively long delay.

Given that laypeople are generally aware of the negative effects of longer delays on memory performance, why did long-delay participants show overconfidence? Guided by Koriat's (1997) cue-utilisation theory, we suggested that the relationship between retrieval fluency and accuracy may break down over time, and that this may lead to poorer calibration when witnesses are questioned after a relatively long delay if people continue to rely on retrieval fluency to guide their confidence judgements. Consistent with this idea, our exploratory analyses revealed that response times were significantly related to accuracy in the immediate-delay condition but not in the short- or long-delay conditions. In contrast, response times were significantly related to confidence regardless of delay. In other words, people gave higher confidence judgements to fast responses than to relatively slow responses, even when response times did not provide a good indicator of accuracy. These findings may explain why long-delay participants showed overconfidence. Specifically, long-delay participants may have interpreted relatively high retrieval fluency as a sign that responses were likely to be correct and, as such, gave higher confidence judgements than were justified by the accuracy of those responses. This explanation is consistent with previous research showing that the confidence-accuracy relationship breaks down when retrieval fluency provides a misleading cue to accuracy. For example, Experiment 2 and other published research shows that when people are exposed to post-event misinformation, this information tends to come to mind more quickly and, as such, is sometimes reported with higher confidence than details originating from the initial event, producing overconfidence (Flowe et al., 2019).

To date, relatively few studies have looked at the fluency-accuracy relationship after a relatively long delay (i.e., 1 week or more). Existing work suggests that the retrieval fluency-accuracy relationship can be maintained when witnesses are asked to identify a perpetrator from a line-up 1 week after the crime (Sauerland et al., 2019; Sauerland & Sporer, 2009). To our knowledge, however, the current study is the first to examine the effect of retention interval on the relationship between retrieval fluency and eyewitness recall accuracy after a long delay. It is important to note that our measure of response time was relatively imprecise, so the correlations we have reported may underestimate the actual strength of the retrieval fluency-accuracy relationship in our study. This might explain why the response

time-accuracy correlation also failed to reach significance in the short-delay condition.

Consistent with previous research, we found that people's beliefs about their general memory ability did not significantly predict their confidence when they recalled details shortly after witnessing a mock crime event (Saraiva et al., 2020). We predicted that self-rated memory ability would have a larger impact on confidence when the delay was long compared to when the delay was relatively short. We found that people gave lower confidence judgements after longer delays, yet their beliefs about their general memory ability did not affect this pattern. One explanation for these findings is that the metamemory questionnaires focused on participants' beliefs about their everyday memory ability, and these beliefs may not extend to eyewitness scenarios. Developing new measures that specifically focus on eyewitness recall may help to provide further insight on the impact of self-rated memory ability on witness's confidence judgements. Another explanation is that people may recognise that they have little experience of reporting details in an eyewitness context and therefore rely more on theory-based cues arising from the witnessing or testing situation (e.g., lighting conditions), or experience-based cues that arise from the process of remembering (e.g., retrieval fluency, Koriat, 1997).

In our study, long-delay participants were questioned 1 month after witnessing a mock crime event but in real cases delays may often exceed 10 weeks (Pike et al., 2002). This raises important questions about how very long delays affect the confidence-accuracy relationship. For example, do increasingly longer delays further impair the confidence-accuracy relationship and lead to greater overconfidence? Future research should examine the confidence-accuracy relationship over a wider range of retention intervals to better estimate when and how delay impairs the confidence-accuracy relationship. Understanding how delay affects the confidence-accuracy relationship also has important practical and theoretical implications. For example, it may allow researchers to advise legal decision makers on when after the event confidence no longer provides a reliable indicator of accuracy. Furthermore, investigating how retrieval cues such as retrieval fluency change over a delay may help to provide insight into the cognitive processes underpinning accuracy and confidence decisions.

Finally, future research could examine how long delays affect the specificity of witnesses' responses. We know that general responses (e.g., "dark") tend to be more accurate than specific responses (e.g., "dark blue," Sauer & Hope, 2016), yet witnesses rarely give general responses in their open-ended reports (Brewer et al., 2018). Thus, one possibility is that instructing witnesses to vary the specificity of their responses may help to maintain accuracy over long delays. However, it is important to note that general responses tend to produce poorer calibration and greater underconfidence than specific responses (as per Experiment 3) so encouraging witnesses to report more general responses may enhance memory but at the cost of impairing the confidence-accuracy relationship.

To conclude, Experiment 4 reveals three key findings. First, longer delays appear to reduce eyewitness confidence and impair the confidence-accuracy relationship. Second, how people feel about their general memory ability appears to have little or no impact on how confident they are in their reports. Third, people may rely, at least partly, on retrieval fluency to provide confidence decisions, even when the retrieval fluency-accuracy relationship breaks down. This mechanism may account for why confidence-accuracy calibration is reasonably strong after short delays but is impaired following longer delays. Given that legal decision makers often use confidence as an indicator of accuracy, it is important that eyewitness reports are collected as quickly as possible to maintain the confidence-accuracy relationship.

Part Two

Chapter 7:
Introduction to Cross-Examination

When witnesses or defendants are called to testify in court they will give their evidence (called examination in chief) before the opposing side has the right to cross-examine them on their evidence. The purpose of cross-examination is three-fold. First, cross-examination can be used to elicit evidence in support of one's case. Second, it can be used to cast doubt on or to undermine the witness's evidence and credibility to weaken the opponent's case. Third, it can be used to challenge disputed evidence (Laver, n.d.).

Many legal advocates claim that cross-examination is an essential tool for assessing the accuracy of eyewitness testimony and exposing unreliable witnesses (for a discussion, see Hickey, 1993; Wheatcroft et al., 2004). Yet some scholars have argued that cross-examination is an ineffective method for assessing eyewitness accuracy and only serves to undermine witnesses' credibility (Plotnikoff & Woolfson, 2009; Spencer, 2012). Critics of cross-examination warn that lawyers' questions are often suggestive and confusing, making it difficult for witnesses to respond accurately (Kebbell et al., 2003; Perry et al., 1995; Zajac & Cannan, 2009; Zajac et al., 2003). For example, questions are often phrased to elicit a yes/no response and frequently contain negatives (e.g., "Do you not agree that the man was wearing a blue shirt?") and suppositional statements (e.g., "It's true that the man was wearing a blue shirt, isn't it?").

A growing body of evidence suggests that these cross-examination style questions tend to reduce the accuracy of children and other vulnerable witnesses (e.g., people with intellectual disabilities; Morrison et al., 2019), compared to simpler forms of questioning (Jack & Zajac, 2014; O'Neill & Zajac, 2013; Righarts et al., 2015; Zajac et al., 2003, 2009; Zajac & Hayne, 2003, 2006). One study, for instance, found that 5- and 6-year-old children frequently changed their initial reports under cross-examination, producing a substantial decrease in their accuracy (Zajac & Hayne, 2003). In this study, children visited a local police station and engaged in four activities, such as having their fingerprints taken. Six weeks later, the children were interviewed in a direct examination, where they recalled

everything that they could remember and answered yes/no questions about the event. Eight months after the initial interview, the children watched a video of their direct examination and answered questions about the trip to the police station. The questions were designed to persuade the children to change their response from the initial interview (e.g., "I don't think you really got your photo taken. I think someone told you to say that. That's what really happened, isn't it?"). In total, 85% of children made changes to their initial reports during cross-examination regardless of whether they were accurate, reducing the overall accuracy of their reports. A follow-up study revealed that although 9- and 10-year-old children were less likely to change correct responses than incorrect responses, they still changed over 40% of their initially correct responses, producing a significant reduction in their accuracy (Zajac & Hayne, 2006). Thus, even older children seem to be susceptible to the negative effects of cross-examination.

Concerningly, research suggests that cross-examination can reduce the accuracy of children's reports even when they have a strong memory for the original event. For example, one study found that children gave more accurate reports during direct examination than cross-examination, regardless of the delay between their reports (Righarts et al., 2015). Furthermore, when children were interviewed again 1 week after cross-examination, the accuracy of their reports was similar to that of their original reports. Together, these findings suggest that children often change their initial reports and provide less accurate responses during cross-examination despite maintaining a good memory for the original event.

While much research has examined how cross-examination affects the accuracy of children's reports, studies have rarely examined how cross-examination affects the accuracy of (non-vulnerable) adults' reports, or their confidence in these reports (Henderson, 2015). The few studies that have examined the effect of cross-examination style questions on adult witnesses' reports suggest that adults too frequently change their responses during cross-examination (Jack & Zajac, 2014; Valentine & Maras, 2011). In a recent study, participants watched a video depicting a mock crime from one of two perspectives, then completed a recall questionnaire including free recall instructions and 8 specific questions about the video (e.g., "Was the girl left or right-handed?"; Valentine & Maras, 2011). Participants in the co-witness condition completed the questionnaire with another participant who had

watched the video from another perspective, whereas control participants completed the questionnaire alone. After a 30-minute filler task, participants completed another recall questionnaire alone and their responses served as their statement. Four weeks later, participants were questioned by a trainee barrister who was told that it would benefit their (fictitious) client if the witness changed four critical points in their statement. During cross-examination, 73% of participants changed their response on at least one of the critical questions, producing a 26% reduction in their accuracy.

Other studies have found that cross-examination style questioning can impair adults' ability to respond accurately, even when their memory would otherwise allow them to do so (Chrobak et al., 2021; Jack & Zajac, 2014; Kebbell et al., 2010; Kebbell & Giles, 2000; Kebbell & Johnson, 2000; Wheatcroft & Ellison, 2012). In one study, participants watched a video of an assault and following a filler task they answered either simple (e.g., "Is it true that the woman went into the house?") or cross-examination style (e.g., "Is it not true that the woman did not go into the house?") questions about the video (Kebbell et al., 2010). Participants who answered cross-examination style questions were 41% less accurate than participants who answered simple questions. Similarly, another study found that the decline in accuracy between two interviews was greater when participants answered cross-examination style questions in the second interview than when they answered simple questions (Jack & Zajac, 2014). Although this reduction in accuracy decreased as age increased, the effect was nevertheless replicated across children, adolescents, and adults. Together, these findings make it clear that even adult witnesses may struggle to provide accurate reports when they are subjected to cross-examination.

What factors might influence the accuracy of adult witnesses' memory reports under cross-examination? A review of the cross-examination literature reveals there may be at least two important factors. First, we know that the type of question the witness is asked can influence the accuracy of a witness's report (Chrobak et al., 2021). But the evidence is mixed on exactly which types of cross-examination style questions are most detrimental to witness accuracy. For example, one study found that people provided fewer correct answers when responding to leading questions (e.g., "The young woman who answered the door had long hair, didn't she?") than when responding to simple questions (Wheatcroft & Woods, 2010). Conversely, another study found that questions containing negatives and

double negatives (e.g., “Did the woman not have black hair?”) produced fewer correct responses than simple questions, but leading questions did not impair accuracy when compared to simple questions (Kebbell & Giles, 2000). These findings highlight the possibility that while some question types may impair eyewitness accuracy during cross-examination, others may have little or no effect. The first aim of Experiment 5 was to systematically examine which types of cross-examination style questions are potentially harmful to eyewitness accuracy, and which are not.

Second, how confident the witness feels about their memory of the witnessed event may influence how accurate they are under cross-examination. Witnesses who are not very confident in their memory performance tend to rely more on external sources of information (e.g., information gleaned from others) than people who are highly confident. For example, people are more likely to accept misinformation when they have relatively low self-esteem (Thorley & Kumar, 2017), or when they mistrust their memory compared to when they are optimistic about their memory ability (van Bergen et al., 2010). Furthermore, witnesses who receive negative feedback about their memory performance give lower confidence judgements and are more likely to change their memory reports than those who receive neutral feedback (Henkel, 2017). Therefore, we might expect that witnesses will be more likely to accept the suggestions in cross-examination style questions and therefore change their original responses when they were not initially highly confident in those responses.

There is also good reason to predict that if witnesses rely on how confident they feel about the accuracy of their memory, then cross-examination may (somewhat counterintuitively) serve to enhance the overall accuracy of witnesses’ reports. Experiments 1 to 4, along with other published work, show that the relationship between confidence and accuracy is reasonably strong in witness recall (Paulo et al., 2019; Roberts & Higham, 2002; Sauer & Hope, 2016; Vredeveldt & Sauer, 2015; Wixted et al., 2018). That is, witnesses tend to be more confident in correct responses than in incorrect responses, particularly when questioned without suggestion. It follows, then, that under cross-examination witnesses should be more likely to change responses that they initially answered incorrectly than responses that they initially answered correctly, thus increasing their overall accuracy. This will

only hold, however, for witnesses who are good at discriminating between correct and incorrect responses when they are initially questioned about the witnessed event. Thus, the second aim of Experiment 5 was to examine whether the accuracy of witnesses' responses during cross-examination is affected by their confidence in those reports, and witnesses' ability to distinguish between correct and incorrect responses. We predicted that participants who are good at discriminating between correct and incorrect responses when initially questioned (i.e., those with higher resolution scores) will make better decisions about changing their responses and show higher accuracy under cross-examination than those who are poor at discriminating between correct and incorrect responses when they are initially questioned.

Finally, there is also evidence to suggest that cross-examination style questions can impair adult witnesses' ability to discriminate between correct and incorrect responses, weakening the confidence-accuracy relationship (Kebbell & Giles, 2000; Kebbell & Johnson, 2000; Wheatcroft et al., 2004). One study, for example, found that participants who answered cross-examination style questions were more confident in incorrect responses and produced a weaker confidence-accuracy correlation ($\gamma = .36$) than participants who answered simple questions ($\gamma = .62$; Kebbell & Johnson, 2000). This finding suggests that cross-examination may not only impair the accuracy of witnesses' reports, but also reduce the informativeness of confidence judgements for assessing the accuracy of those reports. Nonetheless, previous studies have rarely examined the confidence-accuracy relationship separately for different question types, so it is unclear which types of question are most harmful to the confidence-accuracy relationship. Furthermore, some previous studies did not include an initial questioning phase, which may help to preserve witnesses' memory and protect them against subsequent misinformation (Gabbert et al., 2012). Thus, the confidence-accuracy relationship observed in these studies may be weaker than the relationship that might be observed in real cases. In Experiment 5, we assessed the confidence-accuracy relationship for different types of cross-examination style questions and examined whether cross-examination style

questions impaired the confidence-accuracy relationship when participants were provided with an initial recall opportunity before cross-examination.

Understanding how witnesses respond to cross-examination style questions has important implications for the criminal justice system. Lawyers typically assume that reliable and co-operative witnesses are not affected by the complex and suggestive questioning techniques used during cross-examination, so it is important to understand whether these assumptions are accurate (Henderson, 2015). Furthermore, informing lawyers about which question types enhance eyewitness accuracy and which do not may enable them to better assess and improve the accuracy of witnesses' memory reports. The current study also has theoretical implications for our understanding of eyewitness memory. For example, examining whether witnesses use confidence to guide their response change decisions will shed light on how witnesses monitor their memory accuracy, and how they decide which details to report when they recall details about a crime. Although research suggests that people use metacognitive monitoring to guide their response change decisions (Jack & Zajac, 2014), the current study is, to our knowledge, the first to examine the role of confidence in witnesses' decisions to change their responses during cross-examination.

In Experiment 5, we examined whether cross-examination style questions enhance eyewitness accuracy, and whether question type and witness confidence play a role. The procedure was similar to that used in Experiment 2. All participants watched a mock crime video and following a filler task they answered simple questions about the event (e.g., "Were the lockers that the perpetrator tried to break into red?") and rated their confidence in each response. After a delay, participants answered a second memory test consisting of both cross-examination style questions and simple questions, or a memory test consisting of only simple questions. The experiment was preregistered, and the numeric data and corresponding R code are available on the Open Science Framework: https://osf.io/tv9pg/.

Chapter 8:
How Do Different Types of Cross-Examination Style Questions Affect Eyewitness Accuracy?

Experiment 5

Method

Participants & Design

Participants were allocated to one of two conditions (cross-examination, control). In the cross-examination phase, cross-examination participants answered 5 types of question (*Simple*, *Negative*, *Double negative*, *Leading*, *Leading-with-feedback*), whereas control participants only answered simple questions. The motivation for this design—and for the approach we took for the power-analysis described below—was two-fold. First, we wanted to examine the effect of question type on memory performance as a within-participants comparison to maximise statistical power (i.e., a within-participants analysis in the cross-examination condition). Second, we wanted to examine whether the mere act of undergoing cross-examination affects how participants respond to simple questions and including the control condition enabled us to do this (i.e., a between-participants analysis of performance on simple questions).

An a priori power analysis for linear regression indicated that approximately 130 participants would be sufficient to detect a medium effect size ($f^2 = .15$) with .95 power ($\alpha = .05$). Guided by this and previous research, we aimed to recruit 150 participants to the cross-examination condition. We aimed to recruit 30 participants to the control condition, so that both conditions produced at least 600 observations for simple questions.

In total, we recruited 233 participants. Of these, 147 were first year psychology students at the University of Warwick who participated in partial fulfilment of course requirements. The remaining 86 participants were recruited through Prolific and received £2.50 upon completing the experiment. We excluded those who did not complete part 2 within 3-5 days of completing part 1 ($n = 21$), failed to comply with the criteria outlined in the experiment ($n = 20$), experienced technical difficulties ($n = 7$), or answered an attention check question incorrectly (n

= 4). The final sample consisted of 181 participants (144 women, 35 men, 2 other/undisclosed, *M* = 27.6 years, *SD* = 13.4, range = 18-67). There were 151 participants in the cross-examination condition and 30 in the control condition. The Department of Psychology Research Ethics Committee at the University of Warwick approved this research.

Materials

We used the "Mischievous Melanie" mock crime sequence which contains 97 still coloured images that depict a woman stealing various items from an academic department on a university campus (Rasor et al., 2021). The sequence is 6 minutes and 28 seconds long, with each image being presented for 4 seconds. The images are accompanied by an audio narrative outlining the story. More information about the stimuli is available at https://osf.io/znj2e/.

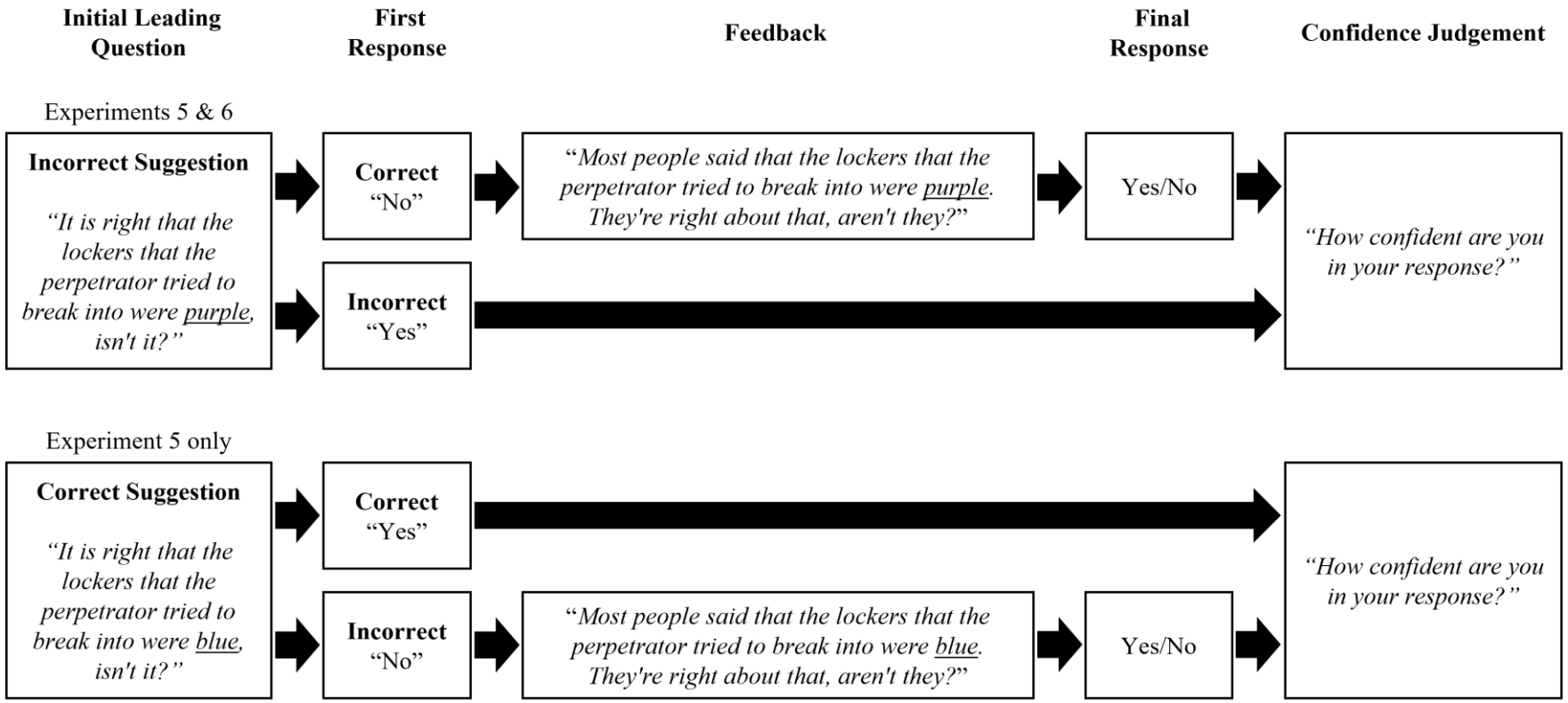
The sequence contains 20 critical items. We created 5 types of question for each critical item:

1. Simple (e.g., "Is it right that the lockers that the perpetrator tried to break into were blue?", correct answer = "yes").
2. Negative: included the word "not" once (e.g., "Is it not right that the lockers that the perpetrator tried to break into were blue?", correct answer = "no").
3. Double negative: included the word "not" twice (e.g., "Is it not right that the lockers that the perpetrator tried to break into were not blue?", correct answer = "yes").
4. Leading: strongly suggested the expected answer (e.g., "It is right that the lockers that the perpetrator tried to break into were purple, isn't it?", correct answer = "no"). Half of leading questions suggested a correct answer (correct answer = "yes"), and half suggested an incorrect answer (correct answer = "no").
5. Leading-with-feedback: A leading question, followed by the suggestion that the correct response to the leading question was "yes" (e.g., "Most people said that the lockers that the perpetrator tried to break into were purple. They're right about that, aren't they?", correct answer = "no"). Feedback was only given when participants answered "no" to the leading question, and only

the participant's final response was included in the analysis. See Figure 8.1 for an overview of leading-with-feedback questions.

Figure 8.1

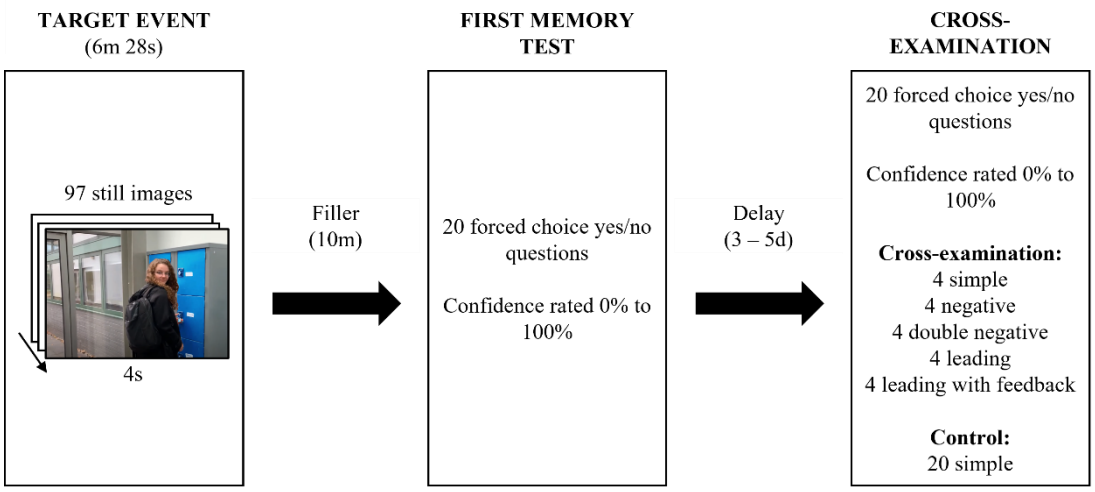
Overview of Leading Questions in Experiments 5 and 6



In the cross-examination condition, questions were divided into 5 sets of 20 questions with each question targeting a different critical item. Each set contained 4 simple, 4 negative, 4 double negative, 4 leading and 4 leading-with-feedback questions. These sets were fully counterbalanced such that each question type was presented the same number of times and the 20 critical items featured equally often in every question type. Control participants always answered the same 20 simple questions. For each set and question type, the correct answer was "yes" to approximately half of the questions and "no" to the other half of the questions. The questions were tested to ensure that participants could not guess the correct answer based on the question format. The counterbalancing scheme is available on the OSF page.

Figure 8.2

The General Procedure for Experiment 5



Procedure

A general overview of the three-phase procedure is presented in Figure 8.2. Participants completed the study online and were randomly assigned to the cross-examination condition or the control condition. Participants were told that the study was about "learning styles and perception of still images" and were asked to comply with several requirements during the experiment (e.g., "Please do not speak to anyone during the experiment"). In Phase 1 (target event), participants watched the

mock crime sequence. The video was followed by two attention check questions and a 10-minute filler task involving logic problems.

In Phase 2, participants completed the first memory test containing 20 forced choice, yes/no questions each referring to a different critical item (e.g., "Were the lockers that the perpetrator tried breaking into blue?"). The questions were presented in chronological order and participants had unlimited time to respond. Participants were told that they must select Yes or No for each response, but that they could elaborate in the textbox next to their chosen response. Participants rated their confidence immediately after each response on an 11-point scale, ranging from 0% (*not at all confident*) to 100% (*very confident*).

Phase 3 began 3-5 days later (M = 3.46, SD = 0.67), and participants completed another 20-item forced choice, yes/no memory test, similar to that in Phase 2 (see Appendix F for an example). Once again, the questions were presented in chronological order and there was no time limit for responding. Cross-examination participants answered 20 questions including 4 of each question type (simple, negative, double negative, leading, and leading-with-feedback), whereas control participants answered 20 simple questions. Participants could elaborate on their responses in the box next to their chosen response and rated their confidence in their responses on the same 11-point scale. Finally, participants were asked if they had complied with the criteria outlined in Phase 1, answered demographic questions, and were debriefed.

Data Coding

The lead researcher coded the responses for both memory tests, blind to question type and condition. For both memory tests, responses were coded as correct if participants selected the correct yes/no response, or if they correctly described the critical item regardless of the specificity of their response. Therefore, if the answer was "light green" then "green" and "lightly coloured" would also be coded as correct. Accuracy was not coded for responses where participants wrote "don't remember" next to their chosen response, because their confidence judgements did not always reflect their lack of certainty (Responses excluded on the first memory test: n = 6, M = 45.00, SD = 44.61, range = 0-100; Responses excluded on the cross-

examination test: n = 57, M = 29.30, SD = 23.44, range = 0-90). Overall, 3,558 observations were coded and included in the analyses.

In line with previous research on cross-examination (Kebbell & Johnson, 2000), the correct yes/no response for negative questions was the opposite to the equivalent simple question, whereas the correct response to double negative questions was the same as the correct response for the equivalent simple question. For example, the correct answer to "Is it not right that the lockers that the perpetrator tried to break into were blue?" would be "no" because the lockers were blue, whereas the correct answer to the simple question "Is it right that the lockers that the perpetrator tried to break into were blue?" or the double negative question "Is it not right that the lockers that the perpetrator tried to break into were not blue?" would be "yes."

For leading-with-feedback questions, only the final answer is included in the analysis. Feedback was only given when participants gave a "no" response to the initial leading question, such that the feedback always supported the same response as the leading question. When no feedback was given (i.e., when participants gave a "yes" response to the leading question), then the question was recoded as a leading question (rather than leading-with-feedback) for the analyses. In total, 857 leading questions and 351 leading-with-feedback questions were included in the analyses.

Participants were coded as changing their responses under cross-examination if they (1) answered correctly on the first memory test and incorrectly on the cross-examination test or (2) answered incorrectly on the first memory test and correctly on the cross-examination test.

Results

Preliminary Analyses

Before conducting our main analyses, we first checked whether performance on simple questions, during the cross-examination phase, was similar between the cross-examination and control conditions. To this end, we calculated mean accuracy on simple questions for each condition. Accuracy was calculated as the number of correct responses divided by the total number of correct and incorrect responses. The non-overlapping CIs indicated that cross-examination participants (M = 72.74, 95%

CI [69.12, 76.36]) were significantly more accurate on simple questions than were control participants (*M* = 64.76, 95% CI [61.36, 68.15]). This result provides preliminary evidence that the act of responding to cross-examination style questions may influence witnesses' responding on a global level. It is possible that cross-examination serves to encourage participants to monitor the accuracy of their memories more carefully as they report them. Consistent with this interpretation, accuracy on the first memory test was similar for the cross-examination (*M* = 76.73, 95% CI [74.65, 78.81]) and control (*M* = 73.49, 95% CI [69.33, 77.64]) conditions, which suggests the differences observed on the simple questions between the two conditions were induced by the cross-examination process. Given that our primary research question focuses on the effect of different types of cross-examination style questions on eyewitness accuracy, our main analyses focus on the cross-examination condition.

Main Analyses

How do different types of cross-examination style questions affect eyewitness accuracy? To test this, we calculated accuracy on the cross-examination test for each question type. The means and CIs for each question type are presented in Table 8.1. Cross-examination participants were significantly more accurate on simple questions than negative and double negative questions. Surprisingly, leading and leading-with-feedback questions produced similar accuracy to simple questions. These findings suggest that negative and double negative questions may impair witnesses' ability to provide an accurate report compared to simple questions, whereas leading and leading-with-feedback questions may have little to no effect on eyewitness accuracy. We return to these findings on memory accuracy in the exploratory analyses.

Table 8.1

Means and 95% Confidence Intervals for Accuracy and Confidence by Question Type in the Cross-Examination Condition in Experiment 5

Question Type	Accuracy		Confidence	
	M	95% CI	*M*	95% CI
Simple	72.74	69.12, 76.36	65.69	62.37, 69.01
Negative	46.85	42.89, 50.82	62.68	59.28, 66.09
Double negative	54.03	49.72, 58.34	61.89	58.41, 65.36
Leading	76.45	73.67, 79.23	66.77	63.57, 69.97
Leading-with-feedback	74.22	69.54, 78.90	62.02	57.84, 66.21

How do cross-examination style questions affect the relationship between eyewitness confidence and accuracy? To test this, we first created calibration curves by plotting participants' memory accuracy during the cross-examination test against their confidence. We used 3 confidence bins (0-40, 50-80, 90-100) which were chosen such that each confidence bin contained at least 100 observations (see Table 8.2 for the count data). Figure 8.3 shows the confidence-accuracy calibration plots for each question type and Table 8.3 shows the calibration statistics. At the lowest level of confidence (0-40) all question types produced a similar, moderate, level of accuracy. At medium (50-80) to high (90-100) levels of confidence, however, negative questions and double negative questions produced significantly lower accuracy than simple, leading, and leading-with-feedback questions. These results suggest that questions containing negatives and double negatives impair eyewitness performance at the highest levels of confidence compared to simple questions, but performance at the lowest level of confidence is similar across question types. Put another way, negative and double negative questions impaired eyewitness accuracy more than other types of cross-examination style questions, and this impairment was largest when participants were moderately or highly confident. One important practical application of this result is that confidence judgements may be less informative about eyewitness accuracy when witnesses answer negative and double negative questions, than when they answer other types of cross-examination style questions.

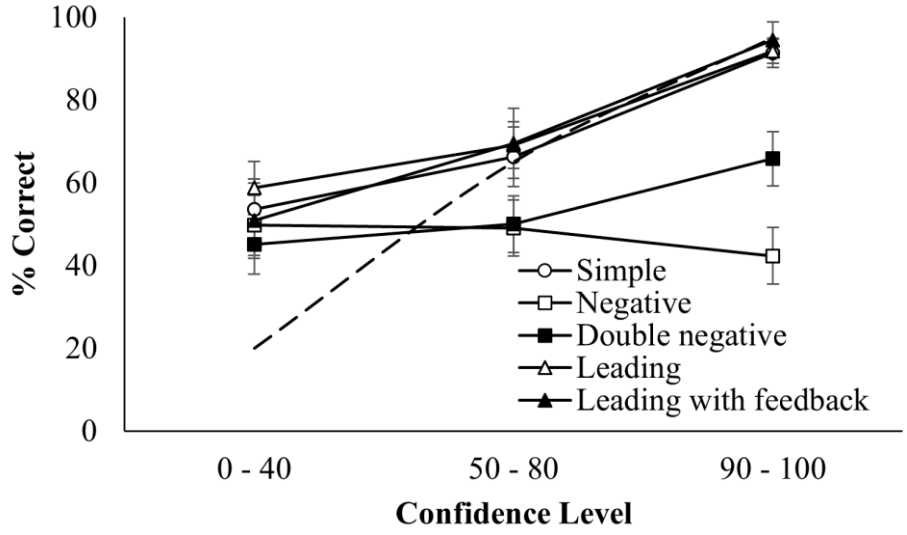
Table 8.2

Count Data for Each Confidence Bin by Cross-Examination Question Type in the Cross-Examination Condition in Experiment 5

Question Type	Confidence Level		
	0 - 40	50 - 80	90 - 100
Simple	179	166	244
Negative	181	208	203
Double negative	186	208	202
Leading	228	256	364
Leading-with-feedback	116	115	110

Figure 8.3

Calibration for Each Cross-Examination Question Type in the Cross-Examination Condition in Experiment 5



Note. The dashed line represents perfect calibration, and the error bars denote the 95% CI around the mean.

Table 8.3

Calibration Statistics and 95% Confidence Intervals by Question Type in the Cross-Examination Condition in Experiment 5

Question Type	*C*	*OU*
Simple	.035 [.019, .050]	-.077 [-.114, -.041]
Negative	.140 [.108, .171]	.154 [.105, .204]
Double negative	.061 [.041, .081]	.078 [.034, .123]
Leading	.041 [.026, .055]	-.090 [-.120, -.060]
Leading-with-feedback	.034 [.013, .055]	-.114 [-.162, -.066]

Note. The *C* statistic reflects the amount of deviation from perfect calibration and ranges from 0 (*perfect calibration*) to 1. The *OU* statistic reflects the amount of over/underconfidence in participants' responses and ranges from -1 (*complete underconfidence*) to 1 (*complete overconfidence*).

Next, we examined whether witnesses were more likely to change their responses during cross-examination when (1) their initial answers were inaccurate than when their initial answers were accurate and (2) they were not very confident in their responses than when they were highly confident in their responses. We conducted a binomial logistic regression on response change (change vs no change) with binary accuracy (correct vs incorrect) and confidence on the first memory test as predictors. Responses were coded as changed when participants' responses changed—in either direction (i.e., accurate to inaccurate or inaccurate to accurate)—between the first memory test and the cross-examination test.

The model revealed that accuracy on the first memory test was a significant predictor of response change, $\chi^2(1) = 42.62$, $p < .001$, and the odds ratio was 1.84, indicating that participants were 84% more likely to change initially inaccurate responses than initially accurate responses. This did not mean, however, that cross-examination improved overall memory accuracy in the current study. It is important to note that although participants changed proportionally more initially incorrect responses (53.10%) than initially correct responses (31.88%), the overall number of correct responses changed ($n = 724$) was greater than the overall number of incorrect responses changed ($n = 369$). Given that participants changed more correct than

incorrect responses, their overall memory accuracy decreased at cross-examination relative to the initial memory test. Of course, had participants performed more poorly on the initial memory test (reporting a much greater proportion of incorrect than correct answers) then we may have observed an increase in overall memory accuracy following cross-examination.

We predicted that participants would be more likely to change responses given initially with low confidence, than responses given initially with high confidence. The model revealed that confidence during the first memory test was a significant predictor of response change under cross-examination, $\chi^2(1) = 107.31$, $p < .001$. The odds ratio was 1.13 indicating that each 10% reduction in confidence increased the likelihood that participants would change their response by 13%. In sum, participants were more likely to change their response when it was inaccurate or when they were not very confident in their response.

Are witnesses who successfully monitor the accuracy of their responses more likely to improve their accuracy during cross-examination? To answer this question, we calculated accuracy change for each participant by subtracting the proportion of correct responses in the cross-examination test from the proportion of correct responses in the initial memory test. Thus, a positive value reflected an increase in accuracy under cross-examination style questioning and a negative value reflected a decrease in accuracy under cross-examination style questioning. We then conducted a multiple linear regression on accuracy change with resolution (*ANRI*) and calibration on the initial memory test and the cross-examination test as predictors. *ANRI* could not be calculated for four participants because they achieved 100% accuracy in the first memory test ($n = 3$) or in the cross-examination test ($n = 1$), so these participants were excluded from the analysis. The remaining 147 cross-examination participants were included in the analysis.

The model was a significant predictor of accuracy change, $F(4, 142) = 3.74$, $p < .01$, $R^2 = .10$. Participants with higher resolution (*ANRI*) scores on the first memory test were more likely to improve their accuracy during the cross-examination phase ($\beta = 0.16$, $p < .01$). None of the other predictors were significant ($p > .05$). Thus, participants who were better able to discriminate between correct

and incorrect responses on the initial test were more likely to improve their overall accuracy under cross-examination.

Exploratory Analyses

Recall that the leading and leading-with-feedback questions suggested the correct response to participants half of the time, and the incorrect response half of the time (and the feedback always implied that the suggestion was correct). Therefore, it is possible that the effect of leading questions on eyewitness accuracy depended on the accuracy of the suggestion. Naturally we might expect witnesses to be more accurate on leading questions that contained a correct suggestion rather than an incorrect suggestion.

To examine this, we calculated the proportion of correct responses by the accuracy of the suggestion in leading questions and leading-with-feedback questions. For leading questions, accuracy was relatively high regardless of whether the question contained a correct (M = 80.24, 95% CI [76.66, 83.83]) or incorrect suggestion (M = 72.57, 95% CI [67.94, 77.20]). Even more surprisingly, participants were more accurate on leading questions that included feedback when those questions contained an incorrect suggestion (M = 89.24, 95% CI [85.01, 93.46]) than when they contained a correct suggestion (M = 27.65, 95% CI [18.44, 36.85]). It is worth noting, however, that participants only received feedback when they initially disagreed with the suggestion (i.e., answered “no”) in the leading question. Given that participants displayed relatively high accuracy on leading questions, it was relatively rare for participants to receive feedback on questions containing a correct suggestion because they usually answered “yes” to the initial question. Out of 341 leading-with-feedback questions containing a correct suggestion, participants only received feedback 29% of the time (n = 99, excluding questions where participants answered “don’t know”). Together, these findings suggest that participants rarely changed their responses after feedback, even when it steered them towards the correct response. This may partly explain why leading-with-feedback questions did not affect eyewitness accuracy during cross-examination style questioning.

Whereas leading questions had little to no effect on eyewitness accuracy, negative and double negative questions significantly impaired the accuracy of witnesses' reports. One explanation for this finding is that participants were confused

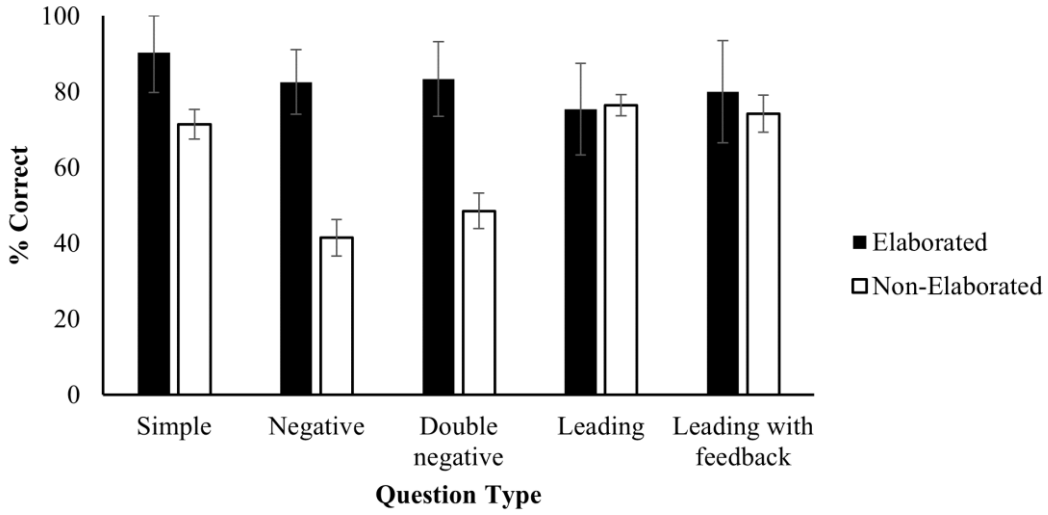
about how to answer negative and double negative questions and, as a result, often gave the wrong answer even when they remembered the critical detail accurately. If participants are confused by negative and double negative questions, it might be expected that they would take longer to comprehend these questions than other types of cross-examination style questions. To examine this, we calculated the mean response time for each question type except leading-with-feedback questions, because these questions consisted of two parts (the initial leading question and the subsequent feedback). To control for unmeasured variables that might affect response times (e.g., lapses in attention), response times more than 3 standard deviations above the mean were removed and not included in the analysis (n = 61, 1.71%). As expected, participants took longer to answer negative (M = 10.23, 95% CI [9.32, 11.14]) and double negative (M = 13.64, 95% CI [12.53, 14.75]) questions than simple (M = 7.64, 95% CI [7.01, 8.27]) and leading (M = 8.11, 95% CI [7.57, 8.65]) questions. These results provide some evidence that participants may have found negative and double negative questions confusing and took longer to comprehend these questions than other types of cross-examination style questions.

If participants answered negative and double negative questions incorrectly, despite having an accurate memory of the critical details in question, then we might also expect them to be more accurate when they elaborated on their responses than when they did not elaborate on their responses. In other words, if witnesses remember a detail accurately then they should be able to describe it correctly, even if they do not know whether "yes" or "no" is the correct response to give. Overall, cross-examination participants only elaborated on 10.89% ($n_{elaborated}$ = 323, $n_{non\text{-}elaborated}$ = 2643) of responses. Figure 8.4 shows that elaborated responses tended to be more accurate than non-elaborated responses. This difference is largest for negative and double negative questions, with a smaller but significant difference for simple questions. There was no significant difference, however, for leading and leading-with-feedback questions. These results suggest that witnesses may have sometimes answered negative and double negative questions incorrectly because they were unsure of the correct response, even when they remembered the critical details accurately. Furthermore, these data suggest that encouraging witnesses to elaborate on their yes/no responses may be important for maximising the accuracy of

eyewitness reports under cross-examination.

Figure 8.4

Mean Accuracy for Elaborated and Non-Elaborated Responses by Question Type in the Cross-Examination Condition in Experiment 5



Note. Error bars denote the 95% CI around the mean.

Conclusion

To summarise, Experiment 5 suggests that the effect of cross-examination style questions on the accuracy of witness memory may depend on question type and witnesses' confidence in their initial responses. Whereas negative and double negative questions impaired participants' ability to provide accurate responses compared to simple questions, leading and leading-with-feedback questions did not significantly affect the accuracy of their responses. Furthermore, participants were more likely to change their initially incorrect and low confidence responses than their initially correct and high confidence responses. Finally, participants who were better at discriminating between correct and incorrect responses on the first memory test were more likely to improve their accuracy under cross-examination.

The finding that leading and leading-with-feedback questions did not impair witness memory accuracy is difficult to reconcile with Experiment 2 and other research showing that, in the face of suggestive questioning techniques, people often

report misinformation and sometimes do so with high confidence (Gabbert et al., 2012; Jack et al., 2014; Loftus et al., 1978; Paterson & Kemp, 2006). Why, then, did leading questions not affect the accuracy of participants' reports in the current study? Recall that cross-examination participants answered 4 questions of each type of cross-examination style question. It is possible that the large number of confusing, cross-examination style questions relative to simple questions (16:4) led participants to believe that the questions were there to trick them into providing the wrong answer. If so, they may have inferred that the suggestions in the leading questions were likely to be incorrect, and only agreed with the suggested response when it matched their memory for the original event. If participants were sceptical about the nature of the cross-examination style questions, then this may also explain why participants rarely changed their responses after they received feedback. Put simply, if participants believed that the purpose of the feedback was to steer them away from the correct response, then it is unsurprising that they often stuck with their original "no" response after feedback.

In Experiment 6, we set out to explore this account by exposing participants to fewer confusing cross-examination style questions. If the large number of confusing questions in Experiment 5 led participants to believe that the questions were designed to trick them, then memory performance on leading questions may be much worse if participants are exposed to fewer confusing cross-examination style questions. Indeed, in real-world legal settings, witnesses might encounter only a small portion of leading questions while under cross-examination so a design in which participants encounter only a few leading questions better mimics the cross-examination scenario.

Chapter 9: Re-Examining The Effect of Cross-Examination Style Questions on Eyewitness Accuracy

The primary aim of Experiment 6 was to examine whether leading questions affect the accuracy of eyewitness reports when the cross-examination test includes fewer confusing questions. To this end, we used the same procedure as that outlined in Experiment 5, but the cross-examination test consisted of 5 cross-examination style questions and 15 filler questions. Additionally, to better simulate the process of direct examination, the first memory test consisted of 20 cued recall questions instead of yes/no questions. Experiment 6 also served as an opportunity to attempt to replicate the key findings around witness accuracy and confidence observed in Experiment 5.

Experiment 6

Method

Participants & Design

In our original pre-registration, we stated that we would collect data from 200 participants. We later realised, however, that this was a miscalculation and that it would not give us enough data points to plot stable calibration curves. Therefore, we continued to collect data until we had at least 250 usable observations per question type, with at least 50 observations in each confidence bin. The changes to our pre-registration are described at https://osf.io/mntdx/.

In total, we recruited 378 participants. Of these, 249 were recruited through Prolific and received £2.20 for completing the experiment. The remaining 129 participants were first year psychology students at the University of Warwick who participated in partial fulfilment of course requirements. We excluded participants who experienced technical difficulties (n = 11), failed to comply with the criteria outlined in the experiment (n = 10) or answered an attention check question incorrectly (n = 5). The final sample consisted of 352 participants (266 women, 79 men, 7 other/undisclosed, M = 33.98, SD = 14.74, range = 17-74). The Department of Psychology Research Ethics Committee at the University of Warwick approved this research. Unlike in Experiment 5, we used a within-participants design and did

not include a separate control group. Thus, all participants answered cross-examination style questions in the cross-examination memory test.

Materials

This experiment used the "Mischievous Melanie" mock crime sequence as in Experiment 5. Of the 20 items in the video, 5 served as critical items and 15 served as filler items. Critical items were chosen such that they were roughly equally distributed throughout the video and memory tests (i.e., items 3, 7, 11, 15 and 19 were selected as critical items). We created 5 questions for each critical item, including one of each type of cross-examination style question (simple, negative, double negative, leading and leading-with-feedback). Filler items were always targeted by simple questions. The questions were created the same way as in Experiment 5, except that leading and leading-with-feedback questions always suggested an incorrect response.

Questions were divided into 5 sets of 20 questions. Each set contained 5 cross-examination style questions each targeting a different critical item, and 15 filler questions. The critical items were counterbalanced so that each item appeared once in each question type, and each question type appeared once in each set. Put simply, participants answered one of each type of cross-examination style question and each targeted a different critical item. The counterbalancing scheme is available on the OSF page for this experiment.

Procedure

Participants completed the three-phase study online and were told that the study was about "learning styles and perception of still images". They were also asked to comply with several criteria during the experiment (e.g., "Please do not speak to anyone during the experiment"). In Phase 1 (target event), participants watched the mock crime sequence, answered two attention check questions, and completed a 10-minute filler task involving logic problems.

In Phase 2 (first memory test), participants completed the first memory test consisting of 20 cued recall questions each referring to a different item in the video (e.g., "What colour were the lockers that the perpetrator tried to break into?") and rated their confidence immediately after each response on an 11-point scale, ranging

from 0% (*not at all confident*) to 100% (*very confident*). As in Experiment 5, questions were presented in chronological order and there was no time limit for participants to give their responses.

Phase 3 (cross-examination) began 3-5 days later (M = 3.35, SD = 0.58), and participants completed a 20-item forced-choice, yes/no memory test which was similar to that used in Experiment 5 (e.g., "Is it right that the lockers that the perpetrator tried to break into were purple?"), except that it consisted of 5 cross-examination style questions and 15 filler questions (see Appendix G for an example memory test). Participants could elaborate on their responses in the box next to their chosen response and they were asked to rate their confidence on the same 11-point scale. Finally, participants indicated if they had complied with the criteria outlined in Phase 1, answered demographic questions, and were debriefed.

Data Coding

The lead researcher coded the responses for both memory tests, blind to question type. For the first memory test, responses were coded as correct if participants correctly described the critical item, regardless of the specificity of their response. Responses in the second memory test were coded in the same way as Experiment 5. Again, responses were coded as correct if participants gave the correct yes/no response or if they correctly described the critical item, regardless of the specificity of their response. Accuracy was not coded for responses where participants wrote "don't know" in either memory test, because their confidence judgements did not always reflect their lack of certainty (Responses excluded on the first memory test: n = 857, M = 30.94, SD = 44.57, range = 0-100; Responses excluded on the second memory test: n = 59, M = 33.05, SD = 35.63, range = 0-100). Overall, 6,124 observations were coded and included in the analyses.

As in Experiment 5, the correct yes/no response to negative questions was the opposite to the equivalent simple question, whereas the correct response to double negative questions was the same as the correct response for the equivalent simple question. In contrast to Experiment 5, however, the correct answer to leading and leading-with-feedback questions was always "no", because the question always suggested an incorrect response.

For leading-with-feedback questions, feedback was only given when participants correctly answered "no" to the initial leading question and only the final, post-feedback answer was included in the analyses. No feedback was given when participants incorrectly answered "yes" to the initial leading question. "No feedback" responses were excluded from the analysis because participants' initial response was always incorrect (i.e., they agreed with the leading suggestion), so recoding them as leading questions (as we did in Experiment 5) would artificially reduce accuracy on leading questions (see deviations from the pre-registration on the OSF page). In total, 66 "no feedback" responses were removed and 255 leading-with-feedback questions were included in the analysis.

Results

Main Analyses

Recall that the main aims were to examine the effect of leading questions when the cross-examination phase included fewer confusing, cross-examination style questions, and to replicate the findings of Experiment 5. The means in Table 9.1 show that, similar to Experiment 5, negative and double negative questions produced the lowest accuracy. However, only negative questions significantly impaired accuracy compared to simple questions. In contrast to Experiment 5, leading and leading-with-feedback questions produced significantly higher accuracy than simple questions. In fact, leading-with-feedback questions produced the highest accuracy of any question type during cross-examination. We return to this finding in the exploratory analyses. These findings suggest that negative questions may impair the accuracy of witnesses' reports compared to other types of cross-examination style questions, whereas leading and leading-with-feedback questions may enhance eyewitness accuracy.

Table 9.1

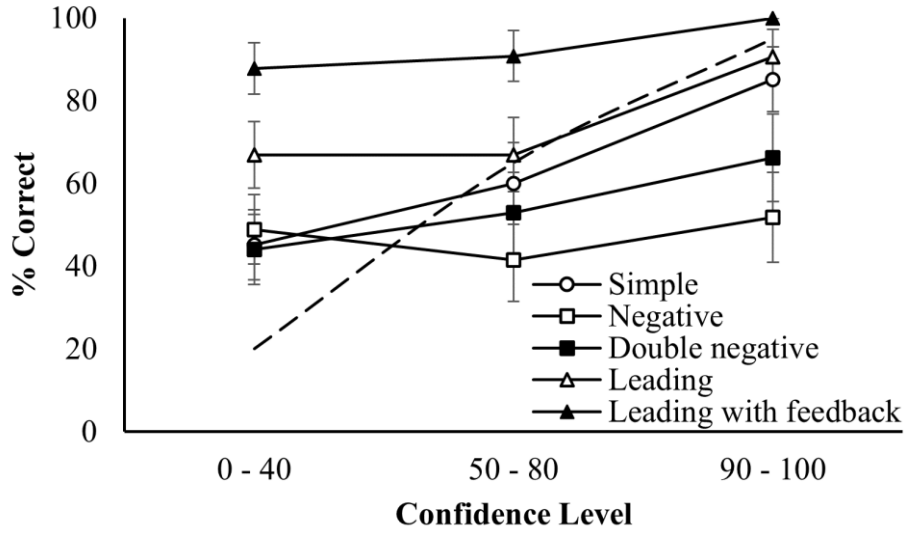
Means and 95% Confidence Intervals for Accuracy and Confidence by Question Type in Experiment 6

Question Type	Accuracy		Confidence	
	M	95% CI	*M*	95% CI
Simple	60.13	54.68, 65.58	52.77	48.89, 56.64
Negative	47.45	41.92, 52.98	51.75	47.91, 55.60
Double negative	52.40	46.85, 57.94	52.36	48.46, 56.27
Leading	72.61	67.67, 77.55	52.58	48.85, 56.31
Leading-with-feedback	91.76	88.38, 95.15	51.18	46.87, 55.48

In Experiment 5, negative and double negative questions impaired eyewitness accuracy at the highest levels of confidence compared to simple questions, but all question types produced a similar level of accuracy at the lowest level of confidence. Figure 9.1 shows that, in contrast to Experiment 5, leading and leading-with-feedback questions produced significantly higher accuracy and greater underconfidence at the lowest level of confidence than other types of cross-examination style questions (see Table 9.2 for the count data and Table 9.3 for the calibration statistics). However, simple, negative and double negative questions produced a similar level of accuracy at low (0-40) to medium (50-80) levels of confidence. Consistent with Experiment 5, negative questions produced the lowest accuracy and greatest overconfidence at the highest level (90-100) of confidence. Double negative questions also produced significant overconfidence, but accuracy did not significantly differ from that for simple or leading questions. Together, these findings suggest that leading and leading-with-feedback questions may produce relatively high accuracy across all levels of confidence, whereas negative and double negative questions may tend to impair accuracy at the highest level of confidence. Given that participants answered fewer confusing questions than in Experiment 5 and still performed relatively well on leading and leading-with-feedback questions, these findings do not support the idea that answering a relatively high number of confusing questions led participants to monitor their memories more carefully than they usually would.

Figure 9.1

Calibration for Each Question Type in the Cross-Examination Test in Experiment 6



Note. The dashed line represents perfect calibration, and the error bars denote the 95% CI around the mean.

Table 9.2

Count Data for Each Confidence Bin by Question Type in the Cross-Examination Test in Experiment 6

Question Type	Confidence Level		
	0 - 40	50 - 80	90 - 100
Simple	135	95	81
Negative	137	94	83
Double negative	134	102	77
Leading	133	106	75
Leading-with-feedback	107	87	61

Table 9.3

Calibration Statistics and 95% Confidence Intervals by Question Type in the Cross-Examination Test in Experiment 6

Question Type	*C*	*OU*
Simple	.036 [.014, .058]	-.074 [-.130, -.017]
Negative	.109 [.071, .147]	.043 [-.024, .110]
Double negative	.060 [.033, .087]	.000 [-.062, .062]
Leading	.100 [.063, .138]	-.200 [-.257, -.144]
Leading-with-feedback	.248 [.196, .301]	-.406 [-.457, -.355]

Note. The *C* statistic reflects the amount of deviation from perfect calibration and ranges from 0 (*perfect calibration*) to 1. The *OU* statistic reflects the amount of over/underconfidence in participants' responses and ranges from -1 (*complete underconfidence*) to 1 (*complete overconfidence*).

Were participants more likely to change initially inaccurate and low confidence responses than initially accurate and high confidence responses? To answer this question, we conducted a binomial logistic regression on response change (change vs no change) with binary accuracy (correct vs incorrect) and confidence on the first memory test as predictors. Again, participants were coded as having changed their responses when the accuracy of their responses changed between the first memory test and the cross-examination test. Consistent with Experiment 5, accuracy on the first memory test, $\chi^2(1) = 41.38$, $p < .001$, was a significant predictor of accuracy change. The odds ratio was 2.33, suggesting that participants were more than twice as likely to change an initially incorrect response than a correct response. Confidence on the first memory test also significantly predicted response change, $\chi^2(1) = 29.33$, $p < .001$, and the odds ratio was 1.10, indicating that each 10% decrease in confidence increased the likelihood that participants would change their responses by 10%. Thus, in line with Experiment 5, participants were more likely to change their initial responses when their responses were not very accurate, or when they were not very confident in their responses.

Are witnesses who are better at discriminating between correct and incorrect responses more likely to improve their accuracy during cross-examination? To test

this, we conducted a multiple linear regression on accuracy change with resolution (*ANRI*) and calibration (*C*) on both the initial memory test and the cross-examination test as predictors. *ANRI* could not be calculated for seven participants because they achieved 100% accuracy in the first memory test ($n = 4$) or in the cross-examination test ($n = 3$). The remaining 345 participants were included in the analyses.

The model significantly predicted accuracy change, F (4, 340) = 7.55, $p <$ .001, $R^2 = .08$. Participants who achieved higher resolution scores in the first memory test, ($\beta = 0.16$, $p < .01$), or relatively strong calibration in the cross-examination memory test, ($\beta = 0.85$, $p < .001$), were more likely to improve their accuracy during cross-examination. Neither calibration on the first memory test or resolution on the cross-examination test significantly predicted accuracy change, $p >$ .05. Thus, in line with Experiment 5, participants who were good at discriminating between accurate and inaccurate responses on the first memory test, or whose confidence judgements were closely aligned with their actual accuracy during the cross-examination phase, were more likely to improve their accuracy during cross-examination.

Exploratory Analyses

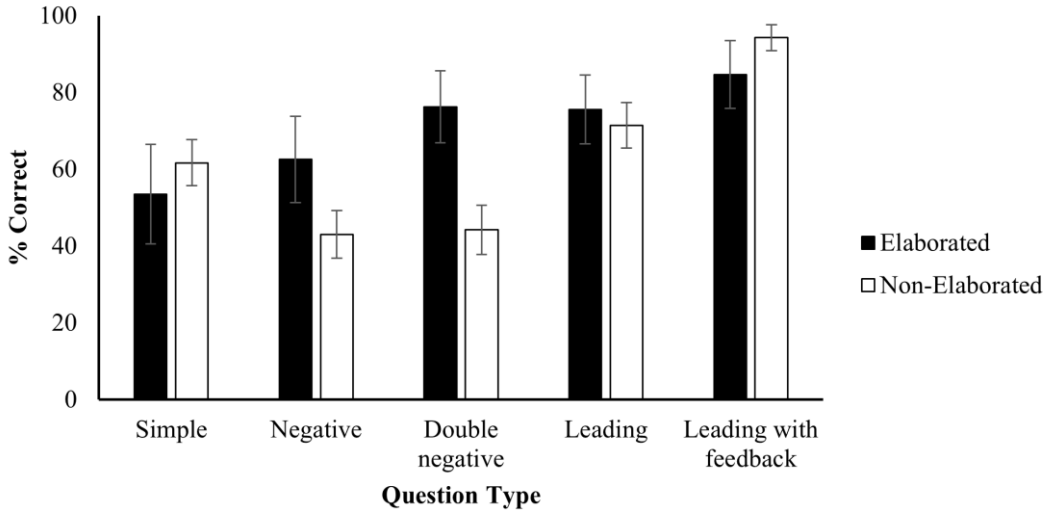
Recall that leading-with-feedback questions produced the highest accuracy and the greatest underconfidence during cross-examination. One possible explanation for these findings is that participants rarely changed their responses after receiving feedback. In this experiment, leading questions always suggested an incorrect response, and participants only received feedback when they disagreed with this suggestion (i.e., when they answered "no" to the preceding leading question). Put another way, participants only received feedback when they answered the initial leading question correctly. One consequence of this is that accuracy will be very high if the feedback fails to convince participants to change their (correct) response to the initial leading question. To examine whether this could explain the high accuracy on leading-with-feedback questions, we calculated how often participants changed their responses to the leading question after feedback. Out of 255 leading-with-feedback questions, participants only changed their responses after feedback 5% ($n = 13$) of the time. These results suggest that participants rarely

changed their responses after feedback, and this may explain why accuracy was particularly high on leading-with-feedback questions.

In Experiment 5, we found that participants were more accurate when they elaborated on their yes/no responses than when they did not elaborate, and this difference was largest for negative and double negative questions. Were participants more accurate on elaborated than non-elaborated responses even when 4 out of 5 types of cross-examination style questions did not impair eyewitness accuracy? Consistent with Experiment 5, Figure 9.2 shows that accuracy on negative and double negative questions was significantly higher when participants elaborated on their responses, than when they did not elaborate on their responses. Accuracy on simple, leading and leading-with-feedback questions was similar regardless of whether participants elaborated on their responses. These results suggest that participants may have sometimes answered negative and double negative questions incorrectly during cross-examination even when their memory should have allowed them to answer correctly.

Figure 9.2

Mean Accuracy for Elaborated and Non-Elaborated Responses by Question Type in the Cross-Examination Test in Experiment 6



Note. Error bars denote the 95% CI around the mean.

Discussion

In Experiments 5 and 6, we examined how different types of cross-examination style questions affect the accuracy of witnesses' reports. We found that negative and double negative questions sometimes impaired the accuracy of witnesses' responses during cross-examination, whereas leading and leading-with-feedback questions did not impair – and sometimes enhanced – the accuracy of witnesses' responses. Participants who were better at discriminating between correct and incorrect responses were more likely to improve the accuracy of their reports during cross-examination. Together, these results suggest that the effect of cross-examination style questions on eyewitness accuracy depends on question type and witnesses' confidence in their responses.

Why did negative questions impair the accuracy of eyewitness reports? One possibility is that participants did not understand the answer that negative questions required and may have adopted different strategies for answering these questions. Recall that the correct response to negative questions was the opposite to the equivalent simple question. It is possible, however, that some participants answered negative questions in the same way as simple questions. For example, if participants believed that the lockers the perpetrator tried to break into were blue, then they may have answered "yes" to the question "Is it not right that the lockers that the perpetrator tried to break into were blue?" regardless of whether it contained the negative (Walker, 1998). If participants answered negative questions in the same way as simple questions, then their responses may have been coded as incorrect even if their memory should have enabled them to respond correctly. This linguistic account may, at least partly, explain why negative questions impaired the accuracy of witnesses' reports compared to simple questions.

Consistent with this interpretation, our exploratory analyses revealed that negative and double negative questions produced similar accuracy to simple questions when participants elaborated on their responses. This finding suggests that participants often provided the incorrect response, even when they could accurately recall the target item. Furthermore, the finding that negative and double negative questions did not impair accuracy when participants elaborated on their responses suggests that encouraging witnesses to elaborate on their responses may provide a

simple technique for maintaining eyewitness accuracy under cross-examination. It is also important to note, however, that linguistic differences across cultures may influence how people respond to cross-examination style questions (Hope et al., 2022). Examining how people from different cultures respond to cross-examination style questions may be a beneficial avenue for future research.

In line with previous work (Kebbell et al., 2010), we found that people tended to take longer to answer negative and double negative questions than simple questions. This finding could indicate that participants struggled to decide on the correct answer to these questions. It is interesting, then, that participants did not acknowledge that their responses to negative and double negative questions could be ambiguous. If participants recognised that their responses were ambiguous, then it might be expected that they would have elaborated on their responses. Alternatively, participants could have lowered their confidence to reflect their lack of certainty that they had chosen the correct answer. In the current research, however, participants rarely elaborated on their responses, and their confidence on negative and double negative questions was similar to that on simple questions. Taken together, these results suggest that participants failed to recognise when their responses were ambiguous on negative and double negative questions. If real witnesses fail to recognise when their responses could be misinterpreted, then they may confidently give a response which does not accurately reflect the content of their memory.

Experiment 2 and other previous research shows that participants who are exposed to misinformation often come to report this information in their memory reports (Flowe et al., 2019; Gabbert et al., 2004; Greene et al., 2021; Ito et al., 2019; Stark et al., 2010). Why, then, did the leading questions in the current study fail to impair the accuracy of eyewitness reports when those questions suggested an incorrect response? One explanation is that participants' original memories were strengthened and protected—at least to some extent—by the recall task they completed before being exposed to misinformation during the cross-examination phase (Gabbert et al., 2012, 2015). In a typical misinformation experiment, mock witnesses are exposed to misinformation before their memory of the critical event is tested and thus are more likely to fall prey to the misinformation. This interpretation is supported by research showing that people who complete a recall task before encountering misinformation are less likely to report misinformation in their later

memory reports than those who do not complete an initial recall task (Gabbert et al., 2012; Geiselman et al., 1986; Memon et al., 2010). This mechanism may also explain why participants rarely changed their responses following negative feedback. Indeed, research suggests that feedback is likely to have a greater impact on eyewitness memory reports when a witness's memory is relatively weak than when their memory is relatively strong (Charman et al., 2010; Iida et al., 2020). Thus, completing an initial memory test may have enabled our participants to perform relatively well on leading and leading-with-feedback questions.

Consistent with a growing body of research on eyewitness confidence and accuracy (Ackerman & Goldsmith, 2008; Weber & Brewer, 2008; Wixted et al., 2018), we found that people were more likely to change responses that they were not very confident in than responses that they were highly confident in. Furthermore, participants who were better at discriminating between correct and incorrect (initial) responses were more likely to improve their accuracy, later, during cross-examination. These findings are consistent with previous research showing that adult witnesses use metacognitive monitoring to guide their responses during cross-examination (Jack & Zajac, 2014). Together, our findings and previous research suggest that questions that enable witnesses to better estimate their memory accuracy may help to improve the accuracy of their responses during cross-examination. Put simply, if lawyers ask questions that enhance witnesses' ability to discriminate between correct and incorrect responses, then witnesses should be more likely to change inaccurate responses than accurate responses, increasing the accuracy of their reports. If, however, lawyers' questions make it hard for witnesses to discriminate between correct and incorrect responses, then witnesses may change their responses indiscriminately at a cost to their overall accuracy.

Our findings highlight at least two fruitful areas for future research. First, it is not clear whether cross-examination has long-term effects on eyewitness memory. We do not yet know whether cross-examination affects the content of witnesses' memory, or which types of cross-examination style questions are the most likely to damage the accuracy of witnesses' future memory reports. Second, as far as we are aware, no published research into cross-examination and witness behaviour has explored the effect of cross-examination style questions on non-cooperative, reluctant or deceptive witnesses. Future research should explore the extent to which

cross-examination style questions influence the responses of witnesses who are not forthcoming with the truth.

To conclude, Experiments 5 and 6 provide the first systematic examination of how different types of cross-examination style questions affect eyewitness accuracy and confidence. To date, there is a lack of guidance and empirical research on how to enhance the accuracy of cooperative witnesses during cross-examination, and little is known about how confidence affects witnesses' responses to cross-examination style questions. For cross-examination to be useful in assessing the credibility of eyewitness evidence—as many legal scholars claim it is—legal decision makers must have a clear and evidence-based understanding of how cross-examination style questions influence a witness's ability to recognise and alter their inaccurate responses. Moreover, it is important to understand whether a witness's expression of confidence truly reflects the accuracy of their memory report when under cross-examination. Our findings are an initial step in advancing understanding on the influence of cross-examination style questions on eyewitness accuracy and confidence. Future research should determine whether these findings generalize to materials and procedures that better mimic the real-world legal process of cross-examination.

Chapter 10:
General Discussion

The aim of this thesis was to improve our understanding of the confidence-accuracy relationship when witnesses report details about a crime. More specifically, the aim of Part One was to examine when confidence judgements are a useful indicator of the accuracy of eyewitness reports during police interviews. The aim of Part Two was to explore the relationship between confidence and accuracy in the context of cross-examination. The main findings of each study will now be summarised, before turning to the practical and theoretical implications of this research.

Summary

In Chapters 4 and 5, we examined how the timing of witnesses' confidence judgements affects the relationship between the accuracy of witnesses' reports and their confidence in those reports. In line with previous research on eyewitness reports (Robinson & Johnson, 1996), Experiments 1-3 showed that the confidence-accuracy relationship was similar regardless of whether confidence judgements were taken immediately after each response (immediate-confidence judgement), or at the end of the memory test (delayed-confidence judgement). Furthermore, Experiment 3 showed that the timing of confidence judgements did not affect the number of details that witnesses reported. Taken together, these findings suggest that both immediate and delayed confidence judgements can provide a useful indicator of accuracy and can be collected without compromising the completeness of witnesses' reports.

In Chapter 4, we also examined two other factors that might impair the confidence-accuracy relationship: the visibility of the witnessed event and exposure to misinformation. Experiment 1 revealed that the visibility of the crime had little to no effect on the confidence-accuracy relationship. Specifically, although night visibility impaired the accuracy of witnesses' reports compared to day visibility, the confidence-accuracy relationship was similar regardless of whether participants viewed the crime with day or night visibility. Conversely, in Experiment 2, exposure to misinformation substantially impaired the confidence-accuracy relationship. When participants answered misleading questions about an item, they were less accurate and more overconfident in their later memory reports than participants who had not

answered misleading questions about the item. Together, these findings suggest that the confidence-accuracy relationship can be relatively strong when people realise they have been exposed to factors that influence their memory performance (i.e., visibility), but the confidence-accuracy relationship may break down if witnesses fail to realise when their memory has been contaminated with misinformation.

In Chapter 6, we examined the effect of retention interval and self-rated memory ability on the confidence-accuracy relationship. The results of Experiment 4 showed that how witnesses' felt about their memory ability did not significantly predict their confidence in their memory reports, but retention interval did: participants were less confident and more underconfident when they were tested after 1 week (short-delay) or 1 month (long-delay) than when they were tested immediately (immediate-delay). Although long-delay participants gave relatively few responses with high (80-100%) confidence, they tended to be more overconfident in those responses than short- and immediate-delay participants. In sum, long-delay participants tended to be more conservative in their confidence judgements than immediate- and short-delay participants but this adjustment in confidence was not sufficient to maintain the confidence-accuracy relationship over a relatively long delay.

In Chapters 8 and 9, we examined how different types of cross-examination style questions affected the accuracy of witnesses' reports and their confidence in those reports. In Experiments 5 and 6, negative and double negative questions tended to produce fewer accurate responses and greater overconfidence at the highest level of confidence than simple, leading, and leading-with-feedback questions. Accuracy on negative and double negative questions was significantly improved, however, when participants elaborated on their responses. These findings suggested that participants often answered negative and double negative questions incorrectly, even when their memory should have enabled them to answer accurately. Finally, we investigated whether witnesses' confidence in their memory reports affected their performance during cross-examination. Participants were more likely to change their initial responses when they were incorrect or given with low confidence than when they were correct or given with high confidence. Furthermore, people who were better at discriminating between accurate and inaccurate responses during an initial memory test were more likely to improve the accuracy of their reports during cross-

examination. Thus, people seem to use their confidence to guide their responses, which can sometimes help witnesses to improve the accuracy of their responses during cross-examination.

Practical Implications

The studies presented in this thesis provide some of the first empirical tests of the confidence-accuracy relationship when witnesses report details about a crime. The results have practical implications for legal decision makers (e.g., lawyers and jurors) who collect eyewitness reports and who must assess the accuracy of those reports, and policy makers who create guidelines on the collection of eyewitness evidence.

In many countries (e.g., UK, US, Australia), investigative interviewers are instructed to collect a free narrative report from an eyewitness before asking more specific follow-up questions (Ministry of Justice, 2011; National Institute of Justice, 2003; Technical Working Group for Eyewitness Evidence, 1999). These recommendations are based on psychological research showing that free reports and open-ended questions tend to elicit more accurate responses than closed questions (Lamb et al., 2007). Although research on the confidence-accuracy relationship has surged in the last two decades, there is currently no guidance on whether confidence judgements should be collected during eyewitness interviews. Some memory scientists have strongly recommended that investigators use confidence judgements to assess the accuracy of eyewitness evidence (Wixted et al., 2018; Wixted & Wells, 2017). This recommendation is based on a large body of identification research showing that the accuracy of eyewitness evidence tends to increase with increasing levels of confidence when confidence judgements are collected under relatively optimal conditions (Brewer & Wells, 2006; Mickes, 2015; Palmer et al., 2013; Sauer et al., 2010). It was not clear, however, whether these patterns held when witnesses report details about a crime. Consistent with identification research, we found that confidence judgements often provided a useful indicator of the accuracy of witnesses' reports, but this relationship broke down when witnesses were exposed to misinformation (Experiment 2) and when they were questioned a relatively long time (in this case, 1 month) after the crime (Experiment 4). These findings suggest that confidence judgements taken during eyewitness interviews may provide useful

information for investigators when confidence judgements are collected relatively soon after the crime and when witnesses are not exposed to misinformation.

Although our findings provide new insight about the relationship between eyewitness accuracy and confidence in police interviews, it should be noted that the encoding and testing conditions in the current research differed from the conditions normally experienced by real witnesses. For example, our participants typically completed online, written memory tests in a situation that was relatively devoid of social context. Real witnesses, however, are typically interviewed by an investigator who might influence the content of their memory reports and their confidence in those reports. Research suggests that even subtle, non-verbal feedback from an interviewer such as a head nod may be sufficient to inflate witnesses' confidence in their memory reports (Gurney et al., 2014). Furthermore, real witnesses are likely to face social pressures (e.g., to tell investigators what they want to hear and to appear accurate) that are unlikely to be present in online studies (Kovera & Evelo, 2021). It is also worth highlighting that we examined the confidence-accuracy relationship over a relatively limited number of events, but future research should also examine the confidence-accuracy relationship across different stimuli to examine whether our results hold for a broader range of stimuli. In sum, future research should replicate our findings under different encoding and testing conditions to develop a better understanding of the confidence-accuracy relationship that is likely to be observed in real cases.

If our results are replicated in more forensically relevant scenarios (e.g., with in-person interviews), then confidence judgements could help to aid legal decision making and increase the likelihood of a successful investigation. For example, confidence judgements could be used to guide investigators' thinking about which details are likely to be correct and worth further investigation, and whether they should search for more witnesses to the crime. Furthermore, informing jurors about the confidence of witnesses' during their initial reports may help to improve legal decision making. Research shows that eyewitness confidence can become inflated over time such that witnesses may sometimes appear highly confident in court even when they initially expressed relatively low confidence (Shaw, 1996; Shaw & McClure, 1996; Steblay & Dysart, 2016). Given that triers of fact heavily rely on eyewitness confidence to gauge the accuracy of eyewitness testimony, it is important

that they are made aware of how confident the witness was in their initial memory report.

Our results also provide information on when investigators could collect confidence judgements during police interviews. Recall that studies investigating the confidence-accuracy relationship for eyewitness reports have typically collected confidence judgements either immediately after each response, or by reminding participants of their responses and asking them to rate their confidence in those responses at the end of the memory test. Yet it was not clear whether the timing of confidence judgements affected the usefulness of those judgements for predicting eyewitness accuracy. Experiments 1-3 are the first to show that confidence judgements can provide useful information about eyewitness accuracy regardless of whether they are taken immediately after each response (immediate-confidence judgement), or after witnesses have reported everything that they can remember (delayed-confidence judgement). Furthermore, the results of Experiment 3 suggest that immediate-confidence judgements could be collected without compromising the completeness of witnesses' reports. This is useful information for legal decision makers because it suggests that there are at least two ways that confidence judgements can be taken from eyewitnesses.

Although the timing of confidence judgements did not affect the confidence-accuracy relationship, it is worth noting that delayed-confidence judgements may be more practical in real investigative interviews. In contrast to delayed-confidence judgements, immediate-confidence judgements require the interviewer to stop the witness after they report each detail. This may distract the interviewer from the main goal of any eyewitness interview: collecting a complete and accurate report from the witness. If the interviewer is required to decide when to stop the witness, then they may have less time and cognitive resources to think of appropriate follow-up questions (Hanway et al., 2021). As a result, they may ask follow-up questions that target details that the witness has already reported. Therefore, it might be expected that interviewers will ask more inappropriate follow-up questions and elicit less new information from witnesses when immediate-confidence judgements are collected than when delayed-confidence judgements are collected (see further research for additional discussion on how confidence timing might affect interviewers).

Whereas the timing of confidence judgements did not significantly affect the relationship between eyewitness confidence and accuracy, long retention intervals reduced the accuracy of witnesses' memory reports and impaired the confidence-accuracy relationship (Experiment 4). For example, participants who answered questions about the crime after 1 month were less accurate and significantly more overconfident at the highest level of confidence (80-100%), than participants tested immediately or after 1 week. As mentioned in Chapter 1, this is problematic because legal decision makers often use confidence judgements to gauge the accuracy of eyewitness reports and tend to believe that information given with high confidence is accurate (Bradfield & Wells, 2000; Cutler et al., 1988; Grabman et al., 2021; Key et al., 2022; Slane & Dodson, 2022). One solution to prevent legal decision makers over-relying on eyewitness confidence might be to inform them about the effects of retention interval on the confidence accuracy relationship, and to warn them when witnesses have been interviewed after a long delay. However, research shows that, when people assess the accuracy of a witness's report, they tend to over-rely on confidence and disregard factors that might have influenced the witness's report (Cutler et al., 1988). Thus, warning legal decision makers about the effect of long retention intervals on eyewitness reports may not be sufficient to protect against the negative effects of long retention intervals. If judgements are to be used to assess the accuracy of witnesses' reports, then it seems important to know whether investigators can be trained to distinguish between when confidence judgements are informative about eyewitness accuracy and when they are not, with reasonable accuracy. Therefore, future research could examine whether investigators can be trained to take factors that can impair the confidence-accuracy relationship into account when using confidence judgements to assess the accuracy of eyewitness evidence.

In a perfect legal system, all witnesses would be interviewed immediately after the crime to minimise forgetting and the chance of witnesses' memories being contaminated by misinformation. Yet it is important to note that police time and resources are limited (Dodd, 2020) and, as such, witnesses may not be interviewed until weeks or months after the crime (Pike et al., 2002). Therefore, the finding that the confidence-accuracy relationship weakens over long delays suggests that self-report tools such as the Self-Administered Interview (SAI; Gabbert et al., 2009) may

be important for collecting confidence judgements that are informative about eyewitness accuracy. The SAI consists of instructions and questions designed to elicit accurate and detailed reports from witnesses without the need for an interviewer to question the witness. Research shows that the SAI is effective at eliciting accurate and detailed reports from witnesses and reduces forgetting, such that witnesses who complete the SAI report more correct details in their subsequent reports than witnesses who initially give a free report or who are not given an early recall opportunity (Gabbert et al., 2022; Gawrylowicz et al., 2013, 2014; Horry et al., 2021).

There are at least two ways that the SAI might benefit the confidence-accuracy relationship in real cases. First, the SAI could be administered when it is not possible to interview the witness shortly after the crime, enabling investigators to collect confidence judgements that are likely to be informative about the accuracy of witnesses' reports. Second, research suggests that people who complete the SAI are more resistant to misinformation, which can also have detrimental effects on the confidence-accuracy relationship (Gabbert et al., 2012; Gittins et al., 2015). Moreover, given that the SAI can be administered relatively quickly, there would be less time for witnesses to be exposed to misinformation, meaning that confidence judgements would be more likely to be informative about eyewitness accuracy (see future research for further discussion about using the SAI to maintain the confidence-accuracy relationship over long retention intervals).

Our findings highlight the dangers of relying on confidence to assess the accuracy of a witness's report when they have been exposed to misinformation. Various sources of misinformation can affect witnesses' memory for an event, including conversations with another witness (French et al., 2008; Gabbert et al., 2004; Hope et al., 2008; Ito et al., 2019), misleading questions (Greene et al., 2021; Loftus, 1975), and media reports (Gabbert et al., 2012; Paterson & Kemp, 2006). Experiment 2 adds to a growing body of research showing that people often report misinformation in their later memory reports, and sometimes do so with high confidence (Flowe et al., 2019; Horry et al., 2014). When participants were exposed to misinformation about an item before the final memory test, they were ~20% less accurate at almost every level of confidence than participants who had not been exposed to misinformation. This finding suggests that if witnesses are exposed to

misinformation before they are interviewed by police, then their confidence judgements may not be informative about the accuracy of their reports. This is problematic because it may not be possible to tell whether witnesses' memories have been contaminated by post-event information. As such, it may be difficult, if not impossible, for legal decision makers to gauge when confidence judgements provide useful information about the accuracy of witnesses' reports.

Finally, the findings presented in Part Two have practical implications concerning how lawyers question witnesses during cross-examination. As noted in Chapter 7, it is standard practice for cross-examining lawyers to ask a variety of complex questions that contain negatives, double negatives, and pre-suppositional statements (Perry et al., 1995). Whereas lawyers tend to assume that these questioning techniques are effective at assessing and enhancing the accuracy of eyewitness evidence, our results suggest that some types of cross-examination style questions may impair the accuracy of witnesses' reports. In Experiments 5 and 6, participants' responses tended to be less accurate when they answered negative and double negative questions than when they answered other types of cross-examination style questions. Given that the goal of cross-examination is to assess and enhance the accuracy of eyewitness testimony, it may be wise for lawyers to avoid using negative and double negative questions. Another solution might be to ask witnesses to elaborate on their yes/no responses during cross-examination. When participants elaborated on their responses to negative and double negative questions, the accuracy of their responses was similar to that on simple questions. Yet, it was relatively rare that participants elaborated on their responses at all. In Experiment 5, for instance, participants only elaborated on 10.89% of all responses. Taken together, these findings suggest that encouraging witnesses to elaborate on their responses may help to protect witnesses against the potentially negative effects of cross-examination.

It should be highlighted, however, that witnesses may be more likely to elaborate on their responses when they are questioned by another person (rather than when they complete a written memory test). Therefore, it is possible that negative and double negative questions will have a smaller impact on eyewitness accuracy in real cases than our findings suggest. The effect of negative and double negative questions on eyewitness accuracy may also vary across cultures. Research shows that people in individualistic cultures tend to give more detailed memory reports than

people from collectivist cultures (Hope et al., 2022; Ross & Wang, 2010; Wang et al., 2017). It might be expected, then, that people from individualistic cultures will elaborate on their responses more, and therefore show higher accuracy on negative and double negative questions than people from collectivist cultures. Given that we only collected data from relatively individualistic cultures (i.e., UK and US), our results may underestimate the effect of cross-examination style in more collectivist cultures.

In sum, although confidence judgements might provide useful information for legal decision makers in some circumstances, our data show that eyewitness confidence can sometimes be misleading. Specifically, the findings from Part One suggests that both immediate- and delayed-confidence judgements could provide useful information for legal decision makers if they are collected shortly after the event. Furthermore, the research in Part Two suggests that lawyers should avoid using negative and double negative questions during cross-examination and that prompting witnesses to elaborate on their responses may provide some protection against the negative effects of cross-examination style questioning. Together, the studies in this thesis provide useful information for legal decision makers who collect eyewitness reports and assess the accuracy of those reports, and policy makers who create guidelines for collecting eyewitness evidence.

Theoretical Implications

Although the studies reported in this thesis do not directly examine the basis of witnesses' confidence judgements, our findings provide further empirical support for the leading theory on confidence judgements: cue-utilisation theory (Koriat, 1997). To recap, cue-utilisation theory suggests that people use various cues to make their confidence judgements, including information obtained from the process of remembering (experience-based cues) and their beliefs about the factors that affect memory performance (theory-based cues). Although experience- and theory-based cues often provide a useful indicator of memory performance, they can sometimes lead witnesses to believe that they are more or less accurate than they really are, leading the confidence-accuracy relationship to break down.

One experience-based cue that people might use to judge the accuracy of their memories is processing fluency (Alter & Oppenheimer, 2009). According to

cue-utilisation theory, people perceive information that is relatively easy to process as more accurate, and thus provide it with higher confidence judgements, than information that is more difficult to process (Koriat, 1997). Based on this mechanism, we proposed that showing delayed-confidence participants their responses twice would increase the ease with which those responses were processed, leading them to show greater overconfidence than immediate-confidence participants who only saw their responses once (Chapter 3). We found no evidence to support this hypothesis: the confidence-accuracy relationship was similar for immediate-confidence and delayed-confidence judgements (Experiments 1-3). At first glance, these findings seem to be inconsistent with cue-utilisation theory. However, it may be that delayed-confidence participants recognised that processing fluency was not a useful indicator of accuracy because they had seen their responses twice, and this repeated exposure could serve to increase processing fluency without affecting the accuracy of their reports. If participants recognised that processing fluency was not informative about their memory accuracy, then they may have dismissed this cue when making their confidence judgements. Although we were unable to test this mechanism directly, our results may suggest that people can sometimes use their beliefs about memory (i.e., theory-based cues) to recognise when other, experience-based cues are not informative about the accuracy of their memories.

The findings in Part One also provide evidence that people can use theory-based cues to adjust their confidence judgements to account for factors that impair the accuracy of their memory reports. In Experiment 1, we found that the confidence-accuracy relationship was relatively strong regardless of the visibility of the crime, even though witnesses who saw the event with night visibility were less accurate than those who saw the event with day visibility. One explanation for this finding is that participants recognised that their memory accuracy was impaired by night visibility and reduced their confidence judgements to reflect their lower accuracy. This interpretation is consistent with cue-utilisation theory (Koriat, 1997). According to this model, people use their beliefs about how memory works and the factors that affect memory performance to guide their confidence judgements. Put another way, when people realise that they have been exposed to a factor that impairs memory performance, then they can make the appropriate adjustments in confidence, preserving the confidence-accuracy relationship. This view is supported

by previous research, showing that the confidence-accuracy relationship can remain relatively strong even when memory performance is impaired by factors such as divided attention (Carlson et al., 2017; Pezdek et al., 2020; Sauer & Hope, 2016).

Further evidence for cue-utilisation theory comes from the finding that, when participants were unknowingly exposed to misinformation, they failed to adjust their confidence judgements to reflect the accuracy of their memory reports. In line with previous research (Bonham & González-Vallejo, 2009; Flowe et al., 2019), the results of Experiment 2 revealed that participants who were exposed to misinformation through misleading questions tended to be more overconfident in their later memory reports than those who had not been exposed to misinformation. This finding is consistent with the mechanism that the confidence-accuracy relationship tends to break down when people fail to realise that their memory has been affected by a given factor: participants may not have realised that they were exposed to misinformation, so they failed to adjust their confidence judgements to compensate for their lower accuracy (Koriat, 1997; Leippe et al., 2009). It is also possible, however, that experience-based cues contributed to overconfidence on misled items. Given that participants encountered the misinformation after the crime, the misinformation is likely to have come to mind more quickly than information from the original event. According to cue-utilisation theory, this increased retrieval fluency is likely to be interpreted as a sign that the misinformation was accurate, producing high confidence judgements. Although we did not examine the exact basis of participants' confidence judgements in Experiment 2, our results were nonetheless consistent with what we would expect based on cue-utilisation theory.

The results in Experiment 4 provide some insight into the basis of witnesses' confidence judgements. In Experiment 4, participants who answered questions about the crime after 1 month (long delay) gave fewer high confidence judgements than participants who answered questions immediately (immediate delay) or after 1 week (short delay). In line with cue-utilisation theory, it is possible that long-delay participants recognised that their memory had been impaired by the relatively long retention interval, so they decided to be relatively conservative when making their confidence judgements.

Given that people are generally aware of the effect of retention interval on memory performance (Cormia et al., 2020), why did long delays impair the confidence-accuracy relationship and produce overconfidence at the highest level of confidence? As mentioned in Chapter 6, it is possible that participants used retrieval fluency (an experience-based cue) as well as theory-based cues to guide their confidence judgements. Consistent with this interpretation, we found that response time was significantly related to confidence in all delay conditions. Conversely, response times were significantly and negatively related to accuracy after a short retention interval, but not after a relatively long retention interval. Consistent with cue-utilisation theory, these results suggest that people may use the speed with which information comes to mind to judge the accuracy of their memory reports (Koriat, 1997). These results are also consistent with the idea that the confidence-accuracy relationship can break down if people fail to realise that experience-based cues do not provide useful information about the accuracy of their memory reports. As such, our results build on cue-utilisation theory by suggesting that accurate meta-memorial beliefs (theory-based cues) about witnessing and testing conditions may not be sufficient to maintain the confidence-accuracy relationship when people continue to rely on experience-based cues that are not informative about their memory accuracy. Put simply, even when people correctly realise that a factor such as retention interval has affected their memory performance, it is not guaranteed that confidence judgements will provide a useful indicator of eyewitness accuracy.

Finally, the findings in Part Two contribute to a growing body of evidence suggesting that people use confidence to guide their memory output (Evans & Fisher, 2011; Goldsmith et al., 2002; Shapira & Pansky, 2019; Weber & Brewer, 2008). When participants answered cross-examination style questions about the crime, they were more likely to change their initial responses that they were not very confident in than responses that they were highly confident in (Experiments 5-6). Furthermore, witnesses who were better at discriminating between correct and incorrect responses with their confidence judgements were more likely to improve their accuracy during cross-examination by changing responses that they had previously reported with low confidence. Taken together, these findings suggest that witnesses used their confidence to decide which information to report during cross-examination, and which information they should change. This interpretation is

consistent with previous research showing that people monitor their confidence in their responses, and often withhold information that they are not very confident in (Goldsmith et al., 2002; Koriat & Goldsmith, 1996). These findings also build on Koriat's (1997) cue-utilisation theory by suggesting that people can use their meta-memorial beliefs about their past memory performance not only to guide their confidence judgements but also to guide their memory output on a subsequent memory test.

To summarise, the findings in Part One support cue-utilisation theory and provide further insight into how people might judge the accuracy of their memory reports. Experience- and theory-based cues often help people to reach a reasonable estimate of their memory accuracy and adjust their confidence judgements accordingly, but our results suggest that they can also lead to errors if people fail to realise when these cues are not informative about their memory accuracy. Nonetheless, further research is needed to understand the circumstances in which experience- and theory-based cues are likely to provide a reliable indicator of memory accuracy, and to reveal which cues have the largest impact on people's confidence judgements.

Future Research

The studies presented in this thesis have important practical and theoretical implications, but further research would enhance our understanding of how witnesses make confidence judgements and the usefulness of collecting confidence judgements for assessing the accuracy of witnesses' reports in real cases. In Experiments 1-3, we found that the timing of confidence judgements did not affect the confidence-accuracy relationship or the number of details that witnesses reported. There are good reasons, however, to expect that interviewers will elicit fewer details from witnesses when they collect immediate-confidence judgements than when they collect delayed-confidence judgements. Deciding when to interrupt the witness to take a confidence judgement may increase the cognitive load that interviewers experience. As noted in Chapter 5, this is problematic because interviewers who experience higher cognitive load tend to remember fewer details reported by witnesses, which may lead them to ask inappropriate follow-up questions (Hanway et al., 2021). For example, they may ask questions that target

information that the witness has already provided. Therefore, future research could examine whether collecting immediate-confidence judgements impairs interviewers' ability to remember details reported by eyewitnesses compared to collecting delayed-confidence judgements. While over three decades of research has examined how confidence judgements should be collected (Cutler & Penrod, 1989; Mansour, 2020; Nguyen et al., 2018; Robinson & Johnson, 1996; Wixted & Wells, 2017), few studies have examined how interviewers might be affected by the interview practices recommended by memory scientists.

Another important question for future research is whether it is possible to reverse the detrimental effects of misinformation on the confidence-accuracy relationship. In Experiment 2, we found that when witnesses were exposed to misinformation before completing the memory test, they tended to be overconfident in their responses at almost every level of confidence. Previous research suggests that witnesses who are warned that they may have been exposed to misinformation are less likely to report this information and produce more accurate memory reports than witnesses who do not receive a warning (Blank & Launay, 2014; Echterhoff et al., 2005; Oeberst & Blank, 2012; Oeberst et al., 2021). It is possible, for example, that these warnings might lead witnesses to be more conservative in their confidence judgements than participants were in Experiment 2. Thus, it might be expected that misinformation will have a smaller impact on the confidence-accuracy relationship and produce less overconfidence when witnesses receive a warning about misinformation, then when they do not receive a warning. Given that it may be impossible to prevent witnesses from encountering misinformation, future research should examine whether warning witnesses about the potential for misinformation helps to maintain the confidence-accuracy relationship.

Future research could also investigate whether evidence-based interviewing tools (e.g., SAI; Gabbert et al., 2009) can help to maintain the confidence-accuracy relationship when witnesses report details about a crime after a relatively long delay. In Experiment 4, we found that high confidence judgements were less likely to be indicative of high accuracy when participants were tested after a relatively long delay. Previous work suggests that the SAI elicits more accurate details than other interviewing techniques when witnesses are tested 1 week after the crime (Gabbert et al., 2009; Gawrylowicz et al., 2013; Horry et al., 2021). It is not clear, however,

whether this benefit extends to longer retention intervals (e.g., 1 month) or whether the SAI has benefits for the confidence-accuracy relationship as well as the accuracy of witnesses' reports. Future research investigating whether the SAI can help to maintain the confidence-accuracy relationship over relatively long retention intervals may prove fruitful. If the SAI enhances witnesses' ability to monitor the accuracy of their responses after a long delay, then administering the SAI may provide a relatively simple way to maintain the confidence-accuracy relationship in real cases.

The research presented in Part Two extends our understanding of how cross-examination style questions affect the accuracy and confidence of adult witnesses. The results of Experiments 5 and 6 suggest that participants rarely agreed with the suggestions in the leading questions and often persevered with their original responses after receiving negative feedback. However, it is possible that our results underestimate the impact of leading questions and negative feedback in real cases where witnesses are questioned by another person. Research suggests that people are more likely to accept misinformation when it is presented by a social source than a non-social source (Gabbert et al., 2004; Paterson & Kemp, 2006). It is possible, for example, that witnesses might believe that the cross-examining lawyer has some additional information that undermines their testimony and thus change their responses to avoid appearing inaccurate or untrustworthy. As such, it is possible that real witnesses would be more likely to change their responses than participants in our studies, who may not have trusted that the information in the leading questions was accurate. Thus, it may prove advantageous to explore the effect of cross-examination style questions in social and non-social settings to determine whether social influence affects how people respond to cross-examination style questions. Future research could also examine whether the perceived status and knowledge of the interviewer influences the likelihood that witnesses will acquiesce with suggestions during cross-examination.

Another potentially fruitful avenue for future research is to examine whether witnesses who are instructed to elaborate on their responses answer cross-examination style questions more accurately than those who are not instructed to elaborate. In Experiments 5 and 6, we found that participants were significantly more accurate on negative and double negative questions when they voluntarily elaborated on their yes/no responses, than when they did not elaborate on their responses.

Furthermore, when witnesses did elaborate on their responses to negative and double negative questions, their accuracy on those questions was similar to their accuracy on simple questions. Taken together, these results suggest that encouraging witnesses to elaborate on their responses may improve the accuracy of witnesses' responses to negative and double negative questions. Therefore, it may be beneficial to conduct future research that examines whether instructing witnesses to elaborate improves the accuracy of their responses during cross-examination. If witnesses are more accurate when they are instructed to elaborate on their responses, then this may reveal a simple way of maintaining – or even enhancing – the accuracy of witnesses' responses in the courtroom. Future research could also examine the effect of culture on cross-examination performance, to determine whether there are cultural differences in the extent to which people elaborate on their responses and the accuracy of those responses during cross-examination.

Finally, it may prove useful to examine how different types of cross-examination style questions affect legal decision makers' perceptions of eyewitness accuracy. To our knowledge, only two studies have examined peoples' ability to assess the accuracy of adult witnesses' responses during cross-examination and these studies have yielded mixed results. One study found that complex questions impaired people's ability to discriminate between accurate and inaccurate witnesses (Kebbell et al., 2010). Conversely, another study found that cross-examination style questions only impaired jurors' ability to discriminate between accurate and inaccurate responses when they were accompanied by negative feedback (Wheatcroft et al., 2004). These mixed findings raise the possibility that some question types may make it more difficult for jurors to assess the accuracy of eyewitness evidence than others. Therefore, it may be interesting to examine which question types are the most likely to impair jurors' ability to discriminate between accurate and inaccurate witnesses.

Concluding remarks

The aim of Part One was to examine when a witness's confidence judgements provide a useful indicator of the accuracy of their reports. The aim of Part Two was to examine the confidence-accuracy relationship in the context of cross-examination—a topic that has, until now, received little attention from applied

memory researchers. The results presented in Part One suggest that confidence judgements are informative about the accuracy of witness reports under some conditions. Confidence judgements can be misleading, however, when witnesses are exposed to misinformation and when witnesses are questioned following a relatively long delay after the crime. These findings align with Koriat's (1997) cue-utilisation theory and recent research on the confidence-accuracy relationship, which suggests that people can adjust their confidence appropriately for some, common-sense factors, but not for other factors. The results presented in Part Two suggest that the influence of cross-examination style questions on eyewitness accuracy depends on the type of question and witnesses' confidence in their memory reports. This research highlights avenues for future research on how cross-examination effects witnesses' ability to enhance the accuracy of their memory reports. Taken together, the findings presented in this thesis indicate that confidence judgements may provide useful information about the accuracy of witnesses reports in some circumstances. However, high confidence judgements do not guarantee high accuracy, as the confidence-accuracy relationship can break down when witnesses base their confidence judgements on cues that are not informative about their memory accuracy.

References

Ackerman, R., & Goldsmith, M. (2008). Control over grain size in memory reporting--with and without satisfying knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(5), 1224–1245. https://doi.org/10.1037/a0012938

Alter, A. L., & Oppenheimer, D. M. (2009). Uniting the tribes of fluency to form a metacognitive nation. *Personality and Social Psychology Review: An Official Journal of the Society for Personality and Social Psychology, Inc*, *13*(3), 219–235. https://doi.org/10.1177/1088868309341564

Ames, D. R., Rose, P., & Anderson, C. P. (2006). The NPI-16 as a short measure of narcissism. *Journal of Research in Personality*, *40*(4), 440–450. https://doi.org/10.1016/j.jrp.2005.03.002

Bates, D., Maechler, M., Bolker, B. M., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.48550/arXiv.1406.5823

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Berkowitz, S. R., & Frenda, S. J. (2018). Rethinking the confident eyewitness: A reply to Wixted, Mickes, and Fisher. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *13*(3), 336–338. https://doi.org/10.1177/1745691617751883

Berkowitz, S. R., Garrett, B. L., Fenn, K. M., & Loftus, E. F. (2020). Convicting with confidence? Why we should not over-rely on eyewitness confidence. *Memory*, 1–6. https://doi.org/10.1080/09658211.2020.1849308

Blank, H., & Launay, C. (2014). How to protect eyewitness memory against the misinformation effect: A meta-analysis of post-warning studies. *Journal of Applied Research in Memory and Cognition*, *3*(2), 77–88. https://doi.org/10.1016/j.jarmac.2014.03.005

Bonham, A. J., & González-Vallejo, C. (2009). Assessment of calibration for reconstructed eye-witness memories. *Acta Psychologica*, *131*(1), 34–52. https://doi.org/10.1016/j.actpsy.2009.02.008

Bornstein, B. H., & Zickafoose, D. J. (1999). I know I know it, I know I saw it: The stability of the confidence–accuracy relationship across domains. *Journal of Experimental Psychology: Applied, 5*(1), 76–88. https://doi.org/10.1037/1076-898X.5.1.76

Bothwell, R. K., Deffenbacher, K. A., & Brigham, J. C. (1987). Correlation of eyewitness accuracy and confidence: Optimality hypothesis revisited. *Journal of Applied Psychology, 72*(4), 691–695. https://doi.org/10.1037/0021-9010.72.4.691

Bradfield, A. L., & Wells, G. L. (2000). The perceived validity of eyewitness identification testimony: a test of the five Biggers criteria. *Law and Human Behavior, 24*(5), 581–594. https://doi.org/10.1023/a:1005523129437

Bradfield, A. L., Wells, G. L., & Olson, E. A. (2002). The damaging effect of confirming feedback on the relation between eyewitness certainty and identification accuracy. *Journal of Applied Psychology, 87*(1), 112–120. https://doi.org/10.1037/0021-9010.87.1.112

Brewer, N. (2006). Uses and abuses of eyewitness identification confidence. *Legal and Criminological Psychology, 11*(1), 3–23. https://doi.org/10.1348/135532505x79672

Brewer, N., & Burke, A. (2002). Effects of testimonial inconsistencies and eyewitness confidence on mock-juror judgments. *Law and Human Behavior, 26*(3), 353–364. https://doi.org/10.1023/a:1015380522722

Brewer, N., Keast, A., & Sauer, J. D. (2010). Children's eyewitness identification performance: Effects of a Not Sure response option and accuracy motivation. *Legal and Criminological Psychology*, *15*(2), 261–277. https://doi.org/10.1348/135532509X474822

Brewer, N., Vagadia, A. N., Hope, L., & Gabbert, F. (2018). Interviewing witnesses: Eliciting coarse-grain information. *Law and Human Behavior, 42*(5), 458–471. https://doi.org/10.1037/lhb0000294

Brewer, N., Weber, N., & Semmler, C. (2005). Eyewitness Identification. In N. Brewer & K. D. Williams (Eds.), *Psychology and Law: An Empirical Perspective* (pp. 177–221). The Guilford Press.

Brewer, N., & Wells, G. L. (2006). The confidence-accuracy relationship in eyewitness identification: effects of lineup instructions, foil similarity, and

target-absent base rates. *Journal of Experimental Psychology: Applied*, *12*(1), 11–30. https://doi.org/10.1037/1076-898X.12.1.11

Brown, A. S., & Marsh, E. J. (2008). Evoking false beliefs about autobiographical experience. *Psychonomic Bulletin & Review*, *15*(1), 186–190. https://doi.org/10.3758/pbr.15.1.186

Busey, T. A., Tunnicliff, J., Loftus, G. R., & Loftus, E. F. (2000). Accounts of the confidence-accuracy relation in recognition memory. *Psychonomic Bulletin & Review*, *7*(1), 26–48. https://doi.org/10.3758/BF03210724

Cann, D. R., & Katz, A. N. (2005). Habitual acceptance of misinformation: examination of individual differences and source attributions. *Memory & Cognition*, *33*(3), 405–417. https://doi.org/10.3758/bf03193059

Carlson, C. A., Dias, J. L., Weatherford, D. R., & Carlson, M. A. (2017). An Investigation of the Weapon Focus Effect and the Confidence–Accuracy Relationship for Eyewitness Identification. *Journal of Applied Research in Memory and Cognition*, *6*(1), 82–92. https://doi.org/10.1016/j.jarmac.2016.04.001

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. https://doi.org/10.1002/acp.3275

Charman, S. D., Carlucci, M., Vallano, J., & Gregory, A. H. (2010). The selective cue integration framework: a theory of postidentification witness confidence assessment. *Journal of Experimental Psychology: Applied*, *16*(2), 204–218. https://doi.org/10.1037/a0019495

Chrobak, Q. M., Braun, B. E., Smith, A. L., & Zaragoza, M. S. (2021). Can clarifying instructions mitigate the effects of multifaceted questions on susceptibility to suggestion? *Applied Cognitive Psychology*, *35*(6), 1502–1509. https://doi.org/10.1002/acp.3883

Colloff, M. F., Wade, K. A., & Strange, D. (2016). Unfair Lineups Make Witnesses More Likely to Confuse Innocent and Guilty Suspects. *Psychological Science*, *27*(9), 1227–1239. https://doi.org/10.1177/0956797616655789

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

Cormia, A., Shapland, T., Rasheed, A., & Pezdek, K. (2020). Laypeople's beliefs about the effects of common estimator variables on memory. *Memory*, 1–11. Advance online publication. https://doi.org/10.1080/09658211.2020.1868527

Cutler, B. L., & Penrod, S. D. (1989). Forensically relevant moderators of the relation between eyewitness identification accuracy and confidence. *Journal of Applied Psychology*, *74*(4), 650–652. https://doi.org/10.1037/0021-9010.74.4.650

Cutler, B. L., Penrod, S. D., & Stuve, T. E. (1988). Juror decision making in eyewitness identification cases. *Law and Human Behavior*, *12*(1), 41–55. https://doi.org/10.1007/BF01064273

Deffenbacher, K. A. (1980). Eyewitness accuracy and confidence. *Law and Human Behavior*, *4*(4), 243–260. https://doi.org/10.1007/BF01040617

Desmarais, S. L., & Read, J. D. (2011). After 30 years, what do we know about what jurors know? A meta-analytic review of lay knowledge regarding eyewitness factors. *Law and Human Behavior*, *35*(3), 200–210. https://doi.org/10.1007/s10979-010-9232-6

Dodd, V. (2020, July 1). Police in England and Wales facing 'new era of austerity'. *The Guardian*. https://www.theguardian.com/uk-news/2020/jul/01/police-warn-of-cuts-to-funding-even-worse-than-in-austerity-years

Dodson, C. S., & Dobolyi, D. G. (2016). Confidence and eyewitness identifications: The cross-race effect, decision time and accuracy. *Applied Cognitive Psychology*, *30*(1), 113–125. https://doi.org/10.1002/acp.3178

Dodson, C. S., & Krueger, L. E. (2006). I misremember it well: why older adults are unreliable eyewitnesses. *Psychonomic Bulletin & Review*, *13*(5), 770–775. https://doi.org/10.3758/bf03193995

Douglass, A. B., & Steblay, N. (2006). Memory distortion in eyewitnesses: A meta-analysis of the post-identification feedback effect. *Applied Cognitive Psychology*, *20*(7), 859–869. https://doi.org/10.1002/acp.1237

Earhart, B., La Rooy, D. J., Brubacher, S. P., & Lamb, M. E. (2014). An examination of "don't know" responses in forensic interviews with children. *Behavioral Sciences & the Law*, *32*(6), 746–761. https://doi.org/10.1002/bsl.2141

Echterhoff, G., Hirst, W., & Hussy, W. (2005). How eyewitnesses resist misinformation: social postwarnings and the monitoring of memory

characteristics. *Memory & Cognition*, *33*(5), 770–782. https://doi.org/10.3758/bf03193073

Evans, J. R., & Fisher, R. P. (2011). Eyewitness memory: Balancing the accuracy, precision and quantity of information through metacognitive monitoring and control. *Applied Cognitive Psychology*, *25*(3), 501–508. https://doi.org/10.1002/acp.1722

Fisher, R. P., & Geiselman, R. E. (2010). The cognitive interview method of conducting police interviews: Eliciting extensive information and promoting therapeutic jurisprudence. *International Journal of Law and Psychiatry*, *33*(5–6), 321–328. https://doi.org/10.1016/j.ijlp.2010.09.004

Flowe, H. D., Humphries, J. E., Takarangi, M. K., Zelek, K., Karoğlu, N., Gabbert, F., & Hope, L. (2019). An experimental examination of the effects of alcohol consumption and exposure to misleading postevent information on remembering a hypothetical rape scenario. *Applied Cognitive Psychology*, *33*(3), 393–413. https://doi.org/10.1002/acp.3531

Foster, J. L., Huthwaite, T., Yesberg, J. A., Garry, M., & Loftus, E. F. (2012). Repetition, not number of sources, increases both susceptibility to misinformation and confidence in the accuracy of eyewitnesses. *Acta Psychologica*, *139*(2), 320–326. https://doi.org/10.1016/j.actpsy.2011.12.004

Fox, S. G., & Walters, H. A. (1986). The Impact of General Versus Specific Expert Testimony and Eyewitness Confidence Upon Mock Juror Judgment. *Law and Human Behavior*, *10*(3), 215–228. https://doi.org/10.1007/BF01046211

French, L., Garry, M., & Mori, K. (2008). You say tomato? Collaborative remembering leads to more false memories for intimate couples than for strangers. *Memory*, *16*(3), 262–273. https://doi.org/10.1080/09658210701801491

Gabbert, F., Hope, L., Carter, E., Boon, R., & Fisher, R. (2015). The role of initial witness accounts within the investigative process. In G. Oxburgh, T. Myklebust, T. Grant, & R. Milne (Eds.), *Communication in investigative and legal contexts: Integrated approaches from forensic psychology, linguistics and law enforcement* (pp. 107–131). Wiley Blackwell.

Gabbert, F., Hope, L., & Fisher, R. P. (2009). Protecting eyewitness evidence: examining the efficacy of a self-administered interview tool. *Law and Human Behavior*, *33*(4), 298–307. https://doi.org/10.1007/s10979-008-9146-8

Gabbert, F., Hope, L., Fisher, R. P., & Jamieson, K. (2012). Protecting Against Misleading Post-event Information with a Self-Administered Interview. *Applied Cognitive Psychology*, *26*(4), 568–575. https://doi.org/10.1002/acp.2828

Gabbert, F., Hope, L., Horry, R., Drain, T., & Hughes, C. (2022). Examining the efficacy of a digital version of the Self-Administered Interview. *Computers in Human Behavior Reports*, *5*, 100159. https://doi.org/10.1016/j.chbr.2021.100159

Gabbert, F., Memon, A., Allan, K., & Wright, D. B. (2004). Say it to my face: Examining the effects of socially encountered misinformation. *Legal and Criminological Psychology*, *9*(2), 215–227. https://doi.org/10.1348/1355325041719428

Garrett, B. L. (2011). *Convicting the innocent: Where criminal prosecutions go wrong*. Harvard University Press.

Garrett, B. L., Liu, A., Kafadar, K., Yaffe, J., & Dodson, C. S. (2020). Factoring the Role of Eyewitness Evidence in the Courtroom. *Journal of Empirical Legal Studies*, *17*(3), 556–579. https://doi.org/10.1111/jels.12259

Gawrylowicz, J., Memon, A., & Scoboria, A. (2013). Equipping witnesses with transferable skills: the Self-Administered Interview©. *Psychology, Crime & Law*, *20*(4), 315–325. https://doi.org/10.1080/1068316X.2013.777961

Gawrylowicz, J., Memon, A., Scoboria, A., Hope, L., & Gabbert, F. (2014). Enhancing older adults' eyewitness memory for present and future events with the Self-Administered Interview. *Psychology and Aging*, *29*(4), 885–890. https://doi.org/10.1037/a0038048

Geiselman, R. E., Fisher, R. P., Cohen, G., Holland, H. L., & Surtes, L. (1986). Eyewintess responses to leading and misleading questions under the cognitive interview. *Journal of Police Science & Administration*, *14*, 31–39.

Gittins, C. B., Paterson, H. M., & Sharpe, L. (2015). How does immediate recall of a stressful event affect psychological response to it? *Journal of Behavior Therapy and Experimental Psychiatry*, *46*, 19–26. https://doi.org/10.1016/j.jbtep.2014.07.006

Goldsmith, M., Koriat, A., & Weinberg-Eliezer, A. (2002). Strategic regulation of grain size memory reporting. *Journal of Experimental Psychology: General*, *131*(1), 73–95. https://doi.org/10.1037/0096-3445.131.1.73

Grabman, J. H., Cash, D. K., Slane, C. R., & Dodson, C. S. (2021). Improving the interpretation of verbal eyewitness confidence statements by distinguishing perceptions of certainty from those of accuracy. *Journal of Experimental Psychology: Applied.* Advance Online Publication. https://doi.org/10.1037/xap0000362

Greene, C. M., Bradshaw, R., Huston, C., & Murphy, G. (2021). The medium and the message: Comparing the effectiveness of six methods of misinformation delivery in an eyewitness memory paradigm. *Journal of Experimental Psychology: Applied*. https://doi.org/10.1037/xap0000364

Griffin, D., & Tversky, A. (1992). The Weighing of Evidence and the Determinants of Confidence. *Cognitive Psychology*, *24*(3), 411–435. https://doi.org/10.1016/0010-0285(92)90013-R

Gurney, D. J., Vekaria, K. N., & Howlett, N. (2014). A Nod in the Wrong Direction: Does Non-verbal Feedback Affect Eyewitness Confidence in Interviews? *Psychiatry, Psychology and Law*, *21*(2), 241–250. https://doi.org/10.1080/13218719.2013.804388

Gustafsson, P. U., Lindholm, T., & Jönsson, F. U. (2019). Predicting Accuracy in Eyewitness Testimonies With Memory Retrieval Effort and Confidence. *Frontiers in Psychology*, *10*, 703. https://doi.org/10.3389/fpsyg.2019.00703

Gustafsson, P. U., Lindholm, T., & Jönsson, F. U. (2021). Judging the accuracy of eyewitness testimonies using retrieval effort cues. *Applied Cognitive Psychology*, *35*(5), 1224–1235. https://doi.org/10.1002/acp.3854

Hanway, P., Akehurst, L., Vernham, Z., & Hope, L. (2021). The effects of cognitive load during an investigative interviewing task on mock interviewers' recall of information. *Legal and Criminological Psychology*, *26*(1), 25–41. https://doi.org/10.1111/lcrp.12182

Henderson, E. (2015). Bigger fish to fry: Should the reform of cross-examination be expanded beyond vulnerable witnesses? *The International Journal of Evidence & Proof*, *19*(2), 83–99. https://doi.org/10.1177/1365712714568072

Henkel, L. A. (2017). Inconsistencies across repeated eyewitness interviews: supportive negative feedback can make witnesses change their memory reports. *Psychology, Crime & Law*, *23*(2), 97–117. https://doi.org/10.1080/1068316X.2016.1225051

Hickey, L. (1993). Presupposition under cross-examination. *International Journal for the Semiotics of Law - Revue Internationale de Sémiotique Juridique*, *6*(1), 89–109. https://doi.org/10.1007/bf01458741

Hope, L., Anakwah, N., Antfolk, J., Brubacher, S. P., Flowe, H., Gabbert, F., Giebels, E., Kanja, W., Korkman, J., Kyo, A., Naka, M., Otgaar, H., Powell, M. B., Selim, H., Skrifvars, J., Sorkpah, I. K., Sowatey, E. A., Steele, L. C., Stevens, L., … Anonymous. (2022). Urgent issues and prospects at the intersection of culture, memory, and witness interviews: Exploring the challenges for research and practice. *Legal and Criminological Psychology*, *27*(1), 1–31. https://doi.org/10.1111/lcrp.12202

Hope, L., Ost, J., Gabbert, F., Healey, S., & Lenton, E. (2008). '"With a little help from my friends..."': The role of co-witness relationship in susceptibility to misinformation. *Acta Psychologica*, *127*, 476–484. https://doi.org/10.1016/j.actpsy.2007.08.010

Horry, R., Colton, L.-M., & Williamson, P. (2014). Confidence-accuracy resolution in the misinformation paradigm is influenced by the availability of source cues. *Acta Psychologica*, *151*, 164–173. https://doi.org/10.1016/j.actpsy.2014.06.006

Horry, R., Hughes, C., Sharma, A., Gabbert, F., & Hope, L. (2021). A meta-analytic review of the Self-Administered Interview©: Quantity and accuracy of details reported on initial and subsequent retrieval attempts. *Applied Cognitive Psychology*, *35*(2), 428–444. https://doi.org/10.1002/acp.3753

Huff, C. R. (1987). Wrongful Conviction: Societal Tolerance ofInjustice. *Research in Social Problems and Public Policy*, *9*, 99–103.

Iida, R., Itsukushima, Y., & Mah, E. Y. (2021). Detailed information mitigates confidence inflation. *Psychology, Crime & Law*, *27*(7), 704–728. https://doi.org/10.1080/1068316X.2020.1849696

Iida, R., Itsukusima, Y., & Mah, E. Y. (2020). How do we judge our confidence? Differential effects of meta-memory feedback on eyewitness accuracy and confidence. *Applied Cognitive Psychology, 34*(2), 397–408. https://doi.org/10.1002/acp.3625

Innocence Project. (2021). DNA Exonerations in the United States. *Innocence Project*. https://innocenceproject.org/dna-exonerations-in-the-united-states/

Ito, H., Barzykowski, K., Grzesik, M., Gülgöz, S., Gürdere, C., Janssen, S. M. J., Khor, J., Rowthorn, H., Wade, K. A., Luna, K., Albuquerque, P. B., Kumar, D., Singh, A. D., Cecconello, W. W., Cadavid, S., Laird, N. C., Baldassari, M. J., Lindsay, D. S., & Mori, K. (2019). Eyewitness Memory Distortion Following Co-Witness Discussion: A Replication of Garry, French, Kinzett, and Mori (2008) in Ten Countries. *Journal of Applied Research in Memory and Cognition*, *8*(1), 68–77. https://doi.org/10.1016/j.jarmac.2018.09.004

Jack, F., & Zajac, R. (2014). The effect of age and reminders on witnesses' responses to cross-examination-style questioning. *Journal of Applied Research in Memory and Cognition*, *3*(1), 1–6. https://doi.org/10.1016/j.jarmac.2013.12.001

Jack, F., Zydervelt, S., & Zajac, R. (2014). Are co-witnesses special? Comparing the influence of co-witness and interviewer misinformation on eyewitness reports. *Memory*, *22*(3), 243–255. https://doi.org/10.1080/09658211.2013.778291

Jackson, S., & Kleitman, S. (2014). Individual differences in decision-making and the role of cognitive confidence. *Personality and Individual Differences*, *60*, S32. https://doi.org/10.1016/j.paid.2013.07.062

Jaeger, T. F. (2008). Categorical Data Analysis: Away from ANOVAs (transformation or not) and towards Logit Mixed Models. *Journal of Memory and Language*, *59*(4), 434–446. https://doi.org/10.1016/j.jml.2007.11.007

Jores, T., Colloff, M. F., Kloft, L., Smailes, H., & Flowe, H. D. (2019). A meta-analysis of the effects of acute alcohol intoxication on witness recall. *Applied Cognitive Psychology*, *33*(3), 334–343. https://doi.org/10.1002/acp.3533

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. https://doi.org/10.1037/0278-7393.22.5.1304

Kebbell, M., Deprez, S., & Wagstaff, G. (2003). The Direct and Cross-Examination of Complainants and Defendants in Rape Trials: A Quantitative Analysis of Question Type. *Psychology, Crime & Law*, *9*(1), 49–59. https://doi.org/10.1080/10683160308139

Kebbell, M., Evans, L., & Johnson, S. D. (2010). The influence of lawyers' questions on witness accuracy, confidence, and reaction times and on mock jurors' interpretation of witness accuracy. *Journal of Investigative Psychology and Offender Profiling*, *7*(3), 262–272. https://doi.org/10.1002/jip.125

Kebbell, M., & Giles, D. (2000). Some experimental influences of lawyers' complicated questions on eyewitness confidence and accuracy. *The Journal of Psychology*, *134*(2), 129–139. https://doi.org/10.1080/00223980009600855

Kebbell, M., & Johnson, S. D. (2000). Lawyers' questioning: the effect of confusing questions on witness confidence and accuracy. *Law and Human Behavior*, *24*(6), 629–641. https://doi.org/10.1023/a:1005548102819

Kelley, C. M., & Lindsay, D. S. (1993). Remembering mistaken for knowing: Ease of retrieval as a basis for confidence in answers to general knowledge questions. *Journal of Memory and Language, 32*(1), 1–24. https://doi.org/10.1006/jmla.1993.1001

Kenchel, J. M., Loftus, E. F., & Berkowitz, S. R. (2020, March 5). *Eyewitness testimony in actual cases of exoneration*. Annual Conference of the American Psychology-Law Society, New Orleans, Louisiana.

Key, K. N., Neuschatz, J. S., Gronlund, S. D., Deloach, D., Wetmore, S. A., McAdoo, R. M., & McCollum, D. (2022). High eyewitness confidence is always compelling: that's a problem. *Psychology, Crime & Law*, 1–22. https://doi.org/10.1080/1068316X.2021.2007912

Kleitman, S., Hui, J. S.-W., & Jiang, Y. (2019). Confidence to spare: individual differences in cognitive and metacognitive arrogance and competence. *Metacognition and Learning*, *14*(3), 479–508. https://doi.org/10.1007/s11409-019-09210-x

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General*, *126*(4), 349–370. https://doi.org/10.1037/0096-3445.126.4.349

Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review*, *103*(3), 490–517. https://doi.org/10.1037/0033-295x.103.3.490

Koriat, A., & Levy-Sadot, R. (1999). Processes underlying metacognitive judgments: Information-based and experience-based monitoring of one's own knowledge. In S. Chaiken & Y. Trope (Eds.), *Dual-process theories in social psychology* (pp. 483–502). Guilford Press.

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory*, *6*(2), 107–118. https://doi.org/10.1037/0278-7393.6.2.107

Koriat, A., Nussinson, R., Bless, H., & Shaked, N. (2008). Information-based and experience-based metacognitive judgments: Evidence from subjective confidence. In J. Dunlosky & R. A. Bjork (Eds.), *Handbook of metamemory and memory* (pp. 117–135). Psychology Press.

Kovera, M. B., & Evelo, A. J. (2021). Eyewitness identification in its social context. *Journal of Applied Research in Memory and Cognition*, *10*(3), 313–327. https://doi.org/10.1016/j.jarmac.2021.04.003

Lamb, M. E., Orbach, Y., Hershkowitz, I., Horowitz, D., & Abbott, C. B. (2007). Does the type of prompt affect the accuracy of information provided by alleged victims of abuse in forensic interviews? *Applied Cognitive Psychology*, *21*(9), 1117–1130. https://doi.org/10.1002/acp.1318

Laver, N. (n.d.). Cross-examination in criminal cases. *In Brief*. Retrieved 19 March 2022, from https://www.inbrief.co.uk/court-proceedings/cross-examination/

Leippe, M. R. (1980). Effects of integrative memorial and cognitive processes on the correspondence of eyewitness accuracy and confidence. *Law and Human Behavior*, *4*(4), 261–274. https://doi.org/10.1007/BF01040618

Leippe, M. R., & Eisenstadt, D. (2007). Eyewitness confidence and the confidence-accuracy relationship in memory for people. In R. C. L. Lindsay, D. F. Ross, J. D. Read, & M. P. Toglia (Eds.), *The handbook of eyewitness psychology, Vol. 2. Memory for people.* (Vol. 601, pp. 377–425). Lawrence Erlbaum Associates Publishers.

Leippe, M. R., Eisenstadt, D., & Rauch, S. M. (2009). Cueing confidence in eyewitness identifications: influence of biased lineup instructions and pre-identification memory feedback under varying lineup conditions. *Law and Human Behavior*, *33*(3), 194–212. https://doi.org/10.1007/s10979-008-9135-y

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1977). Calibration of probabilities: The state of the art. In H. Jungermann & G. De Zeeuw (Eds.), *Decision Making and Change in Human Affairs. Theory and Decision Library (An International Series in the Philosophy and Methodology of the Social and Behavioral Sciences)* (pp. 275–324). Springer. https://doi.org/10.1007/978-94-010-1276-8_19

Lindholm, T., Jönsson, F. U., & Liuzza, M. T. (2018). Retrieval effort cues predict eyewitness accuracy. *Journal of Experimental Psychology: Applied, 24*(4), 534–542. https://doi.org/10.1037/xap0000175

Lindsay, D. S., & Johnson, M. K. (1989). The eyewitness suggestibility effect and memory for source. *Memory & Cognition, 17*(3), 349–358. https://doi.org/10.3758/bf03198473

Lindsay, D. S., Nilsen, E., & Read, J. D. (2000). Witnessing-condition heterogeneity and witnesses' versus investigators' confidence in the accuracy of witnesses' identification decisions. *Law and Human Behavior, 24*(6), 685–697. https://doi.org/10.1023/a:1005504320565

Lindsay, D. S., Read, J. D., & Sharma, K. (1998). Accuracy and confidence in person identification: The relationship is strong when witnessing conditions vary widely. *Psychological Science, 9*(3), 215–218. https://doi.org/10.1111/1467-9280.00041

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), 79–89. https://doi.org/10.1037/0021-9010.66.1.79

Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime. com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods, 49*(2), 433–442. https://doi.org/10.3758/s13428-016-0727-z

Loftus, E. F. (1975). Leading questions and the eyewitness report. *Cognitive Psychology, 7*(4), 560–572. https://doi.org/10.1016/0010-0285(75)90023-7

Loftus, E. F. (2005). Planting misinformation in the human mind: a 30-year investigation of the malleability of memory. *Learning & Memory, 12*(4), 361–366. https://doi.org/10.1101/lm.94705

Loftus, E. F., Miller, D. G., & Burns, H. J. (1978). Semantic Integration of Verbal Information into a Visual Memory. *Journal of Experimental Psychology: Human Learning and Memory*, *4*(1), 19–31. https://doi.org/10.1037/0278-7393.4.1.19

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. https://doi.org/10.1016/S0022-5371(74)80011-3

Luna, K., & Martín-Luengo, B. (2012). Confidence-Accuracy Calibration with General Knowledge and Eyewitness Memory Cued Recall Questions. *Applied Cognitive Psychology*, *26*(2), 289–295. https://doi.org/10.1002/acp.1822

Mansour, J. K. (2020). The Confidence-Accuracy Relationship Using Scale Versus Other Methods of Assessing Confidence. *Journal of Applied Research in Memory and Cognition*, *9*(2), 215–231. https://doi.org/10.1016/j.jarmac.2020.01.003

Memon, A., Zaragoza, M., Clifford, B. R., & Kidd, L. (2010). Inoculation or antidote? The effects of cognitive interview timing on false memory for forcibly fabricated events. *Law and Human Behavior*, *34*(2), 105–117. https://doi.org/10.1007/s10979-008-9172-6

Michael, R. B., & Garry, M. (2016). Ordered questions bias eyewitnesses and jurors. *Psychonomic Bulletin & Review*, *23*(2), 601–608. https://doi.org/10.3758/s13423-015-0933-1

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. https://doi.org/10.1016/j.jarmac.2015.01.003

Ministry of Justice. (2011). *Achieving Best Evidence in Criminal Proceedings*. https://www.cps.gov.uk/sites/default/files/documents/legal_guidance/best_evidence_in_criminal_proceedings.pdf

Morgan, C. A., 3rd, Southwick, S., Steffian, G., Hazlett, G. A., & Loftus, E. F. (2013). Misinformation can influence memory for recently experienced,

highly stressful events. *International Journal of Law and Psychiatry*, *36*(1), 11–17. https://doi.org/10.1016/j.ijlp.2012.11.002

Morrison, J., Forrester-Jones, R., Bradshaw, J., & Murphy, G. (2019). Communication and cross-examination in court for children and adults with intellectual disabilities: A systematic review. *The International Journal of Evidence & Proof*, *23*(4), 366–398. https://doi.org/10.1177/1365712719851134

National Institute of Justice. (2003). *Eyewitness Evidence: A Trainer's Manual for Law Enforcement*. https://www.ncjrs.gov/nij/eyewitness/188678.pdf

Nguyen, T. B., Abed, E., & Pezdek, K. (2018). Postdictive confidence (but not predictive confidence) predicts eyewitness memory accuracy. *Cognitive Research: Principles and Implications*, *3*(1), 1–13. https://doi.org/10.1186/s41235-018-0125-4

Odinot, G., & Wolters, G. (2006). Repeated recall, retention interval and the accuracy–confidence relation in eyewitness memory. *Applied Cognitive Psychology*, *20*(7), 973–985. https://doi.org/10.1002/acp.1263

Odinot, G., Wolters, G., & Lavender, T. (2009). Repeated partial eyewitness questioning causes confidence inflation but not retrieval-induced forgetting. *Applied Cognitive Psychology*, *23*(1), 90–97. https://doi.org/10.1002/acp.1443

Odinot, G., Wolters, G., & van Giezen, A. (2013). Accuracy, confidence and consistency in repeated recall of events. *Psychology, Crime & Law*, *19*(7), 629–642. https://doi.org/10.1080/1068316X.2012.660152

Oeberst, A., & Blank, H. (2012). Undoing suggestive influence on memory: the reversibility of the eyewitness misinformation effect. *Cognition*, *125*(2), 141–159. https://doi.org/10.1016/j.cognition.2012.07.009

Oeberst, A., Wachendörfer, M. M., Imhoff, R., & Blank, H. (2021). Rich false memories of autobiographical events can be reversed. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(13). https://doi.org/10.1073/pnas.2026447118

O'Neill, S., & Zajac, R. (2013). The role of repeated interviewing in children's responses to cross-examination-style questioning: Cross-examination and repeated interviewing. *British Journal of Psychology*, *104*(1), 14–38. https://doi.org/10.1111/j.2044-8295.2011.02096.x

Ost, J., Easton, S., Hope, L., French, C. C., & Wright, D. B. (2017). Latent variables underlying the memory beliefs of Chartered Clinical Psychologists, Hypnotherapists and undergraduate students. *Memory*, *25*(1), 57–68. https://doi.org/10.1080/09658211.2015.1125927

Palmer, M. A., Brewer, N., Weber, N., & Nagesh, A. (2013). The confidence-accuracy relationship for eyewitness identification decisions: Effects of exposure duration, retention interval, and divided attention. *Journal of Experimental Psychology: Applied*, *19*(1), 55–71. https://doi.org/10.1037/a0031602

Paterson, H. M., & Kemp, R. I. (2006). Comparing methods of encountering post-event information: the power of co-witness suggestion. *Applied Cognitive Psychology*, *20*(8), 1083–1099. https://doi.org/10.1002/acp.1261

Paulo, R. M., Albuquerque, P. B., & Bull, R. (2016). Improving the Enhanced Cognitive Interview With a New Interview Strategy: Category Clustering Recall. *Applied Cognitive Psychology*, *30*(5), 775–784. https://doi.org/10.1002/acp.3253

Paulo, R. M., Albuquerque, P. B., & Bull, R. (2019). Witnesses' Verbal Evaluation of Certainty and Uncertainty During Investigative Interviews: Relationship with Report Accuracy. *Journal of Police and Criminal Psychology*, *34*(4), 341–350. https://doi.org/10.1007/s11896-019-09333-6

Penrod, S., & Cutler, B. (1995). Witness confidence and witness accuracy: Assessing their forensic relation. *Psychology, Public Policy, and Law*, *1*(4), 817–845.

Perry, N. W., McAuliff, B. D., Tam, P., Claycomb, L., Dostal, C., & Flanagan, C. (1995). When lawyers question children: Is justice served? *Law and Human Behavior*, *19*(6), 609–629. https://doi.org/10.1007/bf01499377

Pezdek, K., Abed, E., & Cormia, A. (2021). Elevated stress impairs the accuracy of eyewitness memory but not the confidence–accuracy relationship. *Journal of Experimental Psychology: Applied, 27*(1), 158–169. https://doi.org/10.1037/xap0000316

Pezdek, K., Abed, E., & Reisberg, D. (2020). Marijuana impairs the accuracy of eyewitness memory and the confidence–accuracy relationship Too. *Journal of Applied Research in Memory and Cognition, 9*(1), 60–67. https://doi.org/10.1016/j.jarmac.2019.11.005

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

Plotnikoff, J., & Woolfson, R. (2009). *Measuring up? Evaluating implementation of government commitments to young witnesses in criminal proceedings*. NSPCC and Nuffield Foundation.

Quinlivan, D. S., Neuschatz, J. S., Cutler, B. L., Wells, G. L., McClung, J., & Harker, D. L. (2012). Do pre-admonition suggestions moderate the effect of unbiased lineup instructions? *Legal and Criminological Psychology*, *17*(1), 165–176. https://doi.org/10.1348/135532510X533554

Rasor, S. A., Spearing, E., King-Nyberg, B., Mah, E. Y., Lindsay, D. S., & Wade, K. A. (July, 2021). *Online co-witness discussions can lead to memory conformity. Co-witness conformity in an online paradigm. P*oster presented at the virtual SARMAC Conference 2021.

Righarts, S., Jack, F., Zajac, R., & Hayne, H. (2015). Young children's responses to cross-examination style questioning: the effects of delay and subsequent questioning. *Psychology, Crime & Law*, *21*(3), 274–296. https://doi.org/10.1080/1068316X.2014.951650

Roberts, W. T., & Higham, P. A. (2002). Selecting accurate statements from the cognitive interview using confidence ratings. *Journal of Experimental Psychology: Applied*, *8*(1), 33–43. https://doi.org/10.1037//1076-898x.8.1.33

Robinson, M. D., & Johnson, J. T. (1996). Recall memory, recognition memory, and the eyewitness confidence-accuracy correlation. *The Journal of Applied Psychology*, *81*(5), 587–594. https://doi.org/10.1037/0021-9010.81.5.587

Robinson, M. D., Johnson, J. T., & Herndon, F. (1997). Reaction time and assessments of cognitive effort as predictors of eyewitness memory accuracy and confidence. *The Journal of Applied Psychology*, *82*(3), 416–425. https://doi.org/10.1037/0021-9010.82.3.416

Ross, M., & Wang, Q. (2010). Why We Remember and What We Remember: Culture and Autobiographical Memory. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *5*(4), 401–409. https://doi.org/10.1177/1745691610375555

Saraiva, R. B., Hope, L., Horselenberg, R., Ost, J., Sauer, J. D., & van Koppen, P. J. (2020). Using metamemory measures and memory tests to estimate

eyewitness free recall performance. *Memory*, *28*(1), 94–106. https://doi.org/10.1080/09658211.2019.1688835

Sarwar, F., Allwood, C. M., & Innes-Ker, Å. (2014). Effects of different types of forensic information on eyewitness' memory and confidence accuracy. *The European Journal of Psychology Applied to Legal Context*, *6*(1), 17–27. https://doi.org/10.5093/ejpalc2014a3

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. https://doi.org/10.1007/s10979-009-9192-x

Sauer, J., & Hope, L. (2016). The effects of divided attention at study and reporting procedure on regulation and monitoring for episodic recall. *Acta Psychologica*, *169*, 143–156. https://doi.org/10.1016/j.actpsy.2016.05.015

Sauer, J., Palmer, M. A., & Brewer, N. (2019, June 3). *Pitfalls in using eyewitness confidence to diagnose the accuracy of an individual identification decision*. ResearchGate; American Psychological Association. https://doi.org/10.1037/law0000203

Sauerland, M., Broers, N. J., & Oorsouw, K. (2019). Two field studies on the effects of alcohol on eyewitness identification, confidence, and decision times. *Applied Cognitive Psychology*, *33*(3), 370–385. https://doi.org/10.1002/acp.3493

Sauerland, M., & Sporer, S. L. (2007). Post-decision confidence, decision time, and self-reported decision processes as postdictors of identification accuracy. *Psychology, Crime & Law*, *13*(6), 611–625. https://doi.org/10.1080/10683160701264561

Sauerland, M., & Sporer, S. L. (2009). Fast and confident: postdicting eyewitness identification accuracy in a field study. *Journal of Experimental Psychology: Applied*, *15*(1), 46–62. https://doi.org/10.1037/a0014560

Scheck, B., Neufeld, P. J., & Dwyer, J. (2000). *Actual innocence: Five days to execution and other dispatches from the wrongly convicted*. Doubleday.

Seale-Carlisle, T. M., Grabman, J. H., & Dodson, C. S. (2021). The language of accurate and inaccurate eyewitnesses. *Journal of Experimental Psychology: General*. Advance Online Publication. https://doi.org/10.1037/xge0001152

Semmler, C., Brewer, N., & Wells, G. L. (2004). Effects of postidentification feedback on eyewitness identification and nonidentification confidence. *Journal of Applied Psychology*, *89*(2), 334–346. https://doi.org/10.1037/0021-9010.89.2.334

Semmler, C., Dunn, J., Mickes, L., & Wixted, J. T. (2018). The role of estimator variables in eyewitness identification. *Journal of Experimental Psychology: Applied*, *24*(3), 400–415. https://doi.org/10.1037/xap0000157

Shapira, A. A., & Pansky, A. (2019). Cognitive and metacognitive determinants of eyewitness memory accuracy over time. *Metacognition and Learning*, *14*(3), 437–461. https://doi.org/10.1007/s11409-019-09206-7

Shaw, J. S. (1996). Increases in eyewitness confidence resulting from postevent questioning. *Journal of Experimental Psychology: Applied*, *2*(2), 126–146. https://doi.org/10.1037/1076-898x.2.2.126

Shaw, J. S., & McClure, K. A. (1996). Repeated postevent questioning can lead to elevated levels of eyewitness confidence. *Law and Human Behavior*, *20*(6), 629–653. https://doi.org/10.1007/bf01499235

Singmann, H., Bolker, B., Westfall, J., Aust, F., & Ben-Shachar, M. S. (2018). afex: Analysis of factorial experiments. *R Package*.

Slane, C. R., & Dodson, C. S. (2022). Eyewitness confidence and mock juror decisions of guilt: A meta-analytic review. *Law and Human Behavior*, *46*(1), 45–66. https://doi.org/10.1037/lhb0000481

Spencer, J. R. (2012). Conclusions. In J. R. Spencer & M. E. Lamb (Eds.), *Children and cross-examination: Time to change the rules* (pp. 171–202). Oxford, UK: Hart Publishing.

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. https://doi.org/10.1002/ejsp.2420220205

Sporer, S. L., Penrod, S., Read, D., & Cutler, B. (1995). Choosing, confidence, and accuracy: a meta-analysis of the confidence-accuracy relation in eyewitness identification studies. *Psychological Bulletin, 118*(3), 315–327. https://doi.org/10.1037/0033-2909.118.3.315

Stark, C. E. L., Okado, Y., & Loftus, E. F. (2010). Imaging the reconstruction of true and false memories using sensory reactivation and the misinformation

paradigms. *Learning & Memory*, *17*(10), 485–488. https://doi.org/10.1101/lm.1845710

Steblay, N. K., & Dysart, J. E. (2016). Repeated eyewitness identification procedures with the same suspect. *Journal of Applied Research in Memory and Cognition*, *5*(3), 284–289. https://doi.org/10.1016/j.jarmac.2016.06.010

Technical Working Group for Eyewitness Evidence. (1999). *Eyewitness evidence: A guide for law enforcement*. US Department of Justice, Office of Justice Programs, National Institute of Justice.

Tekin, E., & Roediger, H. L. (2017). The range of confidence scales does not affect the relationship between confidence and accuracy in recognition memory. *Cognitive Research: Principles and Implications*, *2*(1), 1–13. https://doi.org/10.1186/s41235-017-0086-z

Thorley, C., & Kumar, D. (2017). Eyewitness susceptibility to co-witness misinformation is influenced by co-witness confidence and own self-confidence. *Psychology, Crime & Law*, *23*(4), 342–360. https://doi.org/10.1080/1068316X.2016.1258471

Troyer, A. K., & Rich, J. B. (2002). Psychometric properties of a new metamemory questionnaire for older adults. *The Journals of Gerontology. Series B, Psychological Sciences and Social Sciences*, *57*(1), 19–27. https://doi.org/10.1093/geronb/57.1.p19

Tuckey, M. R., & Brewer, N. (2003). The influence of schemas, stimulus ambiguity, and interview schedule on eyewitness memory over time. *Journal of Experimental Psychology: Applied*, *9*(2), 101–118. https://doi.org/10.1037/1076-898x.9.2.101

Unkelbach, C., & Stahl, C. (2009). A multinomial modeling approach to dissociate different components of the truth effect. *Consciousness and Cognition*, *18*(1), 22–38. https://doi.org/10.1016/j.concog.2008.09.006

Valentine, T., & Maras, K. (2011). The effect of cross-examination on the accuracy of adult eyewitness testimony. *Applied Cognitive Psychology*, *25*(4), 554–561. https://doi.org/10.1002/acp.1768

Valentine, T., & Mesout, J. (2009). Eyewitness identification under stress in the London Dungeon. *Applied Cognitive Psychology*, *23*(2), 151–161. https://doi.org/10.1002/acp.1463

van Bergen, S., Horselenberg, R., Merckelbach, H., Jelicic, M., & Beckers, R. (2010). Memory distrust and acceptance of misinformation. *Applied Cognitive Psychology*, *24*(6), 885–896. https://doi.org/10.1002/acp.1595

Vredeveldt, A., & Sauer, J. D. (2015). Effects of eye-closure on confidence-accuracy relations in eyewitness testimony. *Journal of Applied Research in Memory and Cognition*, *4*(1), 51–58. https://doi.org/10.1016/j.jarmac.2014.12.006

Wade, K. A., Nash, R. A., & Lindsay, D. S. (2018). Reasons to Doubt the Reliability of Eyewitness Memory: Commentary on Wixted, Mickes, and Fisher (2018) [Review of *Reasons to Doubt the Reliability of Eyewitness Memory: Commentary on Wixted, Mickes, and Fisher (2018)*]. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *13*(3), 339–342. https://doi.org/10.1177/1745691618758261

Walker, A. G. (1998). Linguistic analysis of two complex competency questions. *27th International Congress of Applied Psychology, San Francisco*.

Wang, Q., Song, Q., & Kim Koh, J. B. (2017). Culture, Memory, and Narrative Self-Making. *Imagination, Cognition and Personality*, *37*(2), 199–223. https://doi.org/10.1177/0276236617733827

Weber, N., & Brewer, N. (2003). The effect of judgment type and confidence scale on confidence-accuracy calibration in face recognition. *The Journal of Applied Psychology*, *88*(3), 490–499. https://doi.org/10.1037/0021-9010.88.3.490

Weber, N., & Brewer, N. (2004). Confidence-accuracy calibration in absolute and relative face recognition judgments. *Journal of Experimental Psychology: Applied*, *10*(3), 156–172. https://doi.org/10.1037/1076-898X.10.3.156

Weber, N., & Brewer, N. (2008). Eyewitness recall: Regulation of grain size and the role of confidence. *Journal of Experimental Psychology: Applied*, *14*(1), 50–60. https://doi.org/10.1037/1076-898X.14.1.50

Wells, G. L., Kovera, M. B., Douglass, A. B., Brewer, N., Meissner, C. A., & Wixted, J. T. (2020). Policy and procedure recommendations for the collection and preservation of eyewitness identification evidence. *Law and Human Behavior*, *44*(1), 3–36. https://doi.org/10.1037/lhb0000359

Wells, G. L., & Murray, D. M. (1984). Eyewitness confidence. In G. L. Wells & E. F. Loftus (Eds.), *Eyewitness testimony: Psychological perspectives.* Cambridge University Press.

Wheatcroft, J. M., & Ellison, L. E. (2012). Evidence in court: witness preparation and cross-examination style effects on adult witness accuracy. *Behavioral Sciences & the Law*, *30*(6), 821–840. https://doi.org/10.1002/bsl.2031

Wheatcroft, J. M., Wagstaff, G. F., & Kebbell, M. R. (2004). The influence of courtroom questioning style on actual and perceived eyewitness confidence and accuracy. *Legal and Criminological Psychology*, *9*(1), 83–101. https://doi.org/10.1348/135532504322776870

Wheatcroft, J. M., & Woods, S. (2010). Effectiveness of Witness Preparation and Cross-Examination Non-Directive and Directive Leading Question Styles on Witness Accuracy and Confidence. *The International Journal of Evidence & Proof*, *14*(3), 187–207. https://doi.org/10.1350/ijep.2010.14.3.353

Wixted, J. T., Mickes, L., & Fisher, R. P. (2018). Rethinking the Reliability of Eyewitness Memory. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *13*(3), 324–335. https://doi.org/10.1177/1745691617734878

Wixted, J. T., & Wells, G. L. (2017). The Relationship Between Eyewitness Confidence and Identification Accuracy: A New Synthesis. *Psychological Science in the Public Interest: A Journal of the American Psychological Society*, *18*(1), 10–65. https://doi.org/10.1177/1529100616686966

Wright, D. B., Self, G., & Justice, C. (2000). Memory conformity: exploring misinformation effects when presented by another person. *British Journal of Psychology*, *91*, 189–202. https://doi.org/10.1348/000712600161781

Zajac, R., & Cannan, P. (2009). Cross-Examination of Sexual Assault Complainants: A Developmental Comparison. *Psychiatry, Psychology and Law*, *16*(sup1), S36–S54. https://doi.org/10.1080/13218710802620448

Zajac, R., Gross, J., & Hayne, H. (2003). Asked and answered: Questioning children in the courtroom. *Psychiatry, Psychology, and Law: An Interdisciplinary Journal of the Australian and New Zealand Association of Psychiatry, Psychology and Law*, *10*(1), 199–209. https://doi.org/10.1375/132187103322300059

Zajac, R., & Hayne, H. (2003). I don't think that's what really happened: the effect of cross-examination on the accuracy of children's reports. *Journal of Experimental Psychology: Applied*, *9*(3), 187–195. https://doi.org/10.1037/1076-898X.9.3.187

Zajac, R., & Hayne, H. (2006). The negative effect of cross-examination style questioning on children's accuracy: older children are not immune. *Applied Cognitive Psychology*, *20*(1), 3–16. https://doi.org/10.1002/acp.1169

Zajac, R., Jury, E., & O'Neill, S. (2009). The role of psychosocial factors in young children's responses to cross-examination style questioning. *Applied Cognitive Psychology*, *23*(7), 918–935. https://doi.org/10.1002/acp.1536

Zaragoza, M. S., & Lane, S. M. (1994). Source misattributions and the suggestibility of eyewitness memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(4), 934–945. https://doi.org/10.1037//0278-7393.20.4.934

Zhu, B., Chen, C., Loftus, E. F., Lin, C., He, Q., Chen, C., Li, H., Xue, G., Lu, Z., & Dong, Q. (2010). Individual differences in false memory from misinformation: cognitive factors. *Memory*, *18*(5), 543–555. https://doi.org/10.1080/09658211.2010.487051

Appendices

Appendix A:
Memory Tests in Experiment 1

Car Theft

1. How many women, if any, were in the video?
2. How many men, if any, were in the video?
3. What was the colour of the coat worn by the thief?
4. What type of footwear was worn by the thief?
5. What was the colour of the gloves worn by the thief?
6. How did the thief hide their face?
7. Which supermarket was the car parked near?
8. What parking zone was the stolen car parked in?
9. What was the colour of the car that the thief looked in before taking the stolen car?
10. What was the colour of the box that the thief rested on?
11. What make was the car that entered the car park shortly before the crime?
12. What was the colour of the car that entered the car park shortly before the crime?
13. What was the victim doing when they left the car?
14. What was the hair colour of the victim?
15. What was the colour of the jacket worn by the victim?
16. Where was the stolen car parked?
17. What was the colour of the stolen car?
18. What was the make of the stolen car?

Mugging

1. How many women, if any, were in the video?
2. How many men, if any, were in the video?
3. What type of building did the victim walk past?
4. What type of footwear was worn by the victim?

5. What was the colour of the jacket worn by the victim?
6. What happened prior to the crime?
7. What was the colour of the hat worn by the witness?
8. What was the colour of the scarf worn by the witness?
9. What did the victim put in their bag?
10. Why did the witness say they needed to leave?
11. How did the thief hide their face?
12. How many people, if any, had facial hair?
13. What was the colour of the stolen bag?
14. What was the colour of the coat worn by the thief?
15. What was the colour of the trousers worn by the thief?
16. What, if anything, did the thief say during the crime?
17. How many people, if any, were wearing gloves?
18. Who, if anyone, fell over during the chase?

Appendix B:
Memory Tests in Experiment 2

Car Theft

1. According to the sign, which parking zone was the stolen car parked in?
2. What colour was the thief's coat?
3. What type of footwear was the thief wearing?
4. What colour was the car that entered the parking zone shortly before the crime?
5. What type of vehicle was parked near the entrance to the parking zone?
6. What did the thief rest on while taking the phone call?
7. Where did the thief put their phone?
8. What did the victim shout when they noticed the theft?
9. How many trolley bays were in the parking zone?
10. What weapon, if any, was the thief carrying
11. What colour was the stolen car?
12. What did the sign at the back of the car park say?

Mugging

1. What number was printed on the building at the start of the video?
2. What type of footwear was the victim wearing?
3. What colour was the hat worn by the witness?
4. What item did the victim put in their bag?
5. Who, if anyone, was wearing gloves?
6. What colour was the stolen bag?
7. Why did the witness say they needed to leave?
8. What, if anything, did the thief say during the crime?
9. What did the thief use to hide their face?
10. What colour were the trousers worn by the thief?
11. What colour was the pattern on the victim's hat?
12. What did the victim and witness do when they were talking before the crime?

Appendix C:
Memory Tests in Experiment 3

Car Theft

1. According to the sign, which parking zone was the stolen car parked in?
2. What colour was the thief's coat?
3. What type of footwear was the thief wearing?
4. What colour was the car that entered the parking zone shortly before the crime?
5. What type of vehicle was parked near the entrance to the parking zone?
6. What did the thief rest on while taking the phone call?
7. Where did the thief put their phone?
8. What did the victim shout when they noticed the theft?
9. How many trolley bays were in the parking zone?
10. What weapon, if any, was the thief carrying
11. What colour was the stolen car?
12. What did the sign at the back of the car park say?

Mugging

1. What number was printed on the building at the start of the video?
2. What type of footwear was the victim wearing?
3. What colour was the hat worn by the witness?
4. What item did the victim put in their bag?
5. Who, if anyone, was wearing gloves?
6. What colour was the stolen bag?
7. Why did the witness say they needed to leave?
8. What, if anything, did the thief say during the crime?
9. What did the thief use to hide their face?
10. What colour were the trousers worn by the thief?
11. What colour was the pattern on the victim's hat?
12. What did the victim and witness do when they were talking before the crime?

Appendix D:
Narcissism Analyses in Experiment 4

To examine whether narcissism influenced people's confidence in their memory performance, we conducted two multiple regressions on mean confidence and over/underconfidence, including Narcissism, MMQ-Satisfaction, MMQ-Ability, and MMQ-Strategy as predictors (Table D.1). The models were significant for confidence, $R^2 = .07$, adjusted $p < .001$, and over/underconfidence, $R^2 = .14$, adjusted $p < .001$. Higher levels of narcissism were associated with confidence judgements and greater over-confidence in the memory test. The interaction between narcissism and MMQ-Satisfaction was also significant. Surprisingly, higher levels of narcissism were associated with increased over-confidence when memory satisfaction was low, but not when memory satisfaction was high. A similar trend was observed with confidence, such that people with higher levels of narcissism and low levels of memory satisfaction tended to give higher confidence judgements, but this effect was only marginally significant following the Benjamini-Hochberg correction.

Table D.1

Outcome of the Multiple Regression Models with Narcissism in Experiment 4

Predictor	Confidence			Over/underconfidence		
	Estimate	SE	Adjusted *p*	Estimate	SE	Adjusted *p*
Narcissism	5.45	0.97	< .001	0.07	0.01	< .001
MMQ Satisfaction	-0.84	1.35	.71	-0.01	0.01	.48
MMQ Ability	-0.28	1.36	.84	-0.01	0.01	.46
MMQ Strategy	0.67	0.98	.71	0.01	0.01	.35
Narcissism x MMQ-Satisfaction	-3.17	1.46	.08	-0.04	0.01	.001
Narcissism x MMQ-Ability	-1.23	1.31	.69	0.00	0.01	.83
Narcissism x MMQ-Strategy	0.42	0.99	.77	-0.01	0.01	.35

Note. p Values corrected with the Benjamini-Hochberg procedure.

Appendix E:
Memory Tests in Experiment 4

Car Theft

1. According to the sign, which parking zone was the stolen car parked in?
2. What colour was the sign for the parking zone that the stolen car was parked in?
3. What colour was the thief's coat?
4. What type of footwear was the thief wearing?
5. What item of clothing did the thief use to hide their face?
6. What colour was the car that entered the parking zone shortly before the crime?
7. What type of vehicle was parked near the entrance to the parking zone?
8. What did the thief rest on while taking the phone call?
9. What colour was the car that the thief looked in before taking the stolen car?
10. Who, if anyone, was wearing gloves?
11. What was the victim doing while walking away from the car?
12. Where did the thief put their phone after taking the phone call?
13. What did the victim shout when they noticed the theft?
14. What weapon, if any, was the thief carrying?
15. What colour was the stolen car?

Mugging

1. What type of footwear was the victim wearing?
2. Who, if anyone, was wearing a hat?
3. What item did the victim put in their bag?
4. What colour was the jacket worn by the victim?
5. Who, if anyone, was wearing gloves?
6. What colour was the stolen bag?
7. What pattern was on the scarf worn by the witness?
8. Why did the witness say they needed to leave?

9. What colour was the pattern on the victim's hat?
10. What, if anything, did the thief say during the crime?
11. What colour were the trousers worn by the victim?
12. Who, if anyone, had facial hair?
13. What colour were the trousers worn by the thief?
14. Who, if anyone, was wearing glasses?
15. Who, if anyone, fell over during the chase?

Appendix F:
Example Cross-Examination Test in Experiment 5

1. Would you say that the time was 11:45 when the perpetrator looked at her phone?
2. Is it not right that the lockers that the perpetrator tried to break into were blue?
3. Is it not true to say that the perpetrator did not take a tablet from the table when she first entered the building?
4. You agree that the letter K was on the sign on the door downstairs, don’t you?
5. It is right to say that there was a cafe sign at the top of the stairs, isn't it?
6. Would it be correct to say that there was a mug on the table when the perpetrator stole the mobile phone?
7. Would you not say that the screen above the elevator button was blue?
8. Is it not right that the letters ERS were not spray painted inside of the elevator?
9. The perpetrator saw a man when she exited the elevator downstairs. It's true that the man was holding a sieve, isn't it?
10. You agree that there was an image of the Eiffel Tower on the poster downstairs, don't you?
11. The perpetrator saw a woman coming downstairs. Would you agree that the woman was wearing a black sweatshirt?
12. Would it not be correct to say that there was a watch on the floor when the perpetrator stopped to tie her shoelace?
13. Would you not say that there was not a 'no smoking' sign on the wall by the drink fountain?
14. It's right that there was a microwave in the kitchen, isn't it?
15. It's correct that there was a plastic fork on the draining board in the kitchen, isn't it?
16. Is it true that there was a tree drawn on the flipchart in the office?
17. Would you not agree that only 1 coat button was visible when the perpetrator placed her coat on the table in the office?

18. Would it not be correct to say that there was not an image of mountains on the laptop screen in the office?
19. You would also say that there was an abstract pattern on the TV screen in the office, wouldn't you?
20. It's right that the perpetrator took a credit card from the desk in the office, isn't it?

Appendix G:
Example Cross-Examination Test in Experiment 6

1. Would you say that the time was 11:45 when the perpetrator looked at her phone?
2. Is it right that the lockers that the perpetrator tried to break into were purple?
3. Is it true that the perpetrator took a laptop from the table when she first entered the building?
4. Do you agree that the letter K was on the sign on the door downstairs?
5. Is it right to say that there was a toilet sign at the top of the stairs?
6. Would it be correct to say that there was a mug on the table when the perpetrator stole the mobile phone?
7. You would say that the screen above the elevator button was red, wouldn't you?
8. Is it right that the letters ERS were spray painted inside of the elevator?
9. The perpetrator saw a man when she exited the elevator downstairs. Is it true that the man was holding a sieve?
10. Do you agree that there was an image of the Leaning Tower of Pisa on the poster downstairs?
11. The perpetrator saw a woman coming downstairs. Would you not agree that the woman was wearing a blue sweatshirt?
12. Would it be correct to say that there was a glove on the floor when the perpetrator stopped to tie her shoelace?
13. Would you say that there was a 'no mobile phones' sign on the wall by the drink fountain?
14. Is it right that there was a toaster in the kitchen?
15. It's correct that there was a plastic fork on the draining board in the kitchen, isn't it? (Feedback if a “no” response was given: Most people said that there was a plastic fork on the draining board. They're right about that, aren't they?).
16. Is it true that there was a tree drawn on the flipchart in the office?
17. Do you agree that 3 coat buttons were visible when the perpetrator placed her coat on the table in the office?
18. Would it be correct to say that there was an image of a beach on the laptop screen in the office?

19. Would you not say that there was not a car advert on the TV screen in the office?
20. Is it right that the perpetrator took a credit card from the desk in the office?