An Exploration of Ways to Improve Metacognitive Monitoring and Maximise the Quantity and Accuracy of Eyewitness Memory Reports

Stacey Taylor-Aldridge

BBehavSci(Psych), BBehavSci(Hons)

School of Psychology

Faculty of Social and Behavioural Sciences

Thesis submitted to Flinders University for award of the degree Doctor of Philosophy

December, 2015

TABLE OF CONTENTS

SUMMARY

Witnesses have difficulty maximising the accuracy of their memory reports about a crime without reducing the amount of information that is provided (i.e., quantity). The research presented in this thesis aimed to make this task easier by exploring whether it was possible to improve people's ability to distinguish between correct and incorrect memories (i.e., monitoring); a factor that is known to impact on the accuracy and quantity of eyewitness memory reports. Specifically, the studies conducted assessed monitoring ability (Type-2 Signal Detection Theory discrimination) in a memory task where participants responded to closed questions that can be answered in just a few words. Study 1 explored whether mnemonic cues outlined in the source monitoring framework could discriminate between correct and incorrect memories. The findings showed that five mnemonic cues predicted response accuracy after controlling for natural monitoring ability (i.e., confidence). Experiments 2 and 3 manipulated knowledge of a selection of these mnemonic cues in an attempt to improve monitoring. The results of these experiments showed that altering knowledge of mnemonic cues does not improve monitoring or have a significant impact on quantity or accuracy. However, the findings also revealed that eyewitnesses rarely withhold information, suggesting that they may have an unwarranted level of trust in their memory. Thus, Experiment 4 involved warning participants about the fallibility of eyewitness memory in addition to manipulating knowledge of mnemonic cues. Specific instructions regarding memory retrieval were also included in Experiment 4 as such instructions have previously been found to improve the way witnesses respond to unanswerable questions. The results showed that none of these manipulations had a significant impact on monitoring, quantity, or accuracy. As the low withholding rates observed across the experiments could have been a consequence of the closed questions that were used, Experiment 5 explored potential differences in retrieval, monitoring, and

control abilities during open-ended and closed questioning. The findings revealed that the low withholding rates observed in Experiments 2-4 were unlikely to have been a consequence of the closed questions that were used. In addition, the results suggested that monitoring is superior during open-ended questioning. However, due to the limited number of closed questions that were included, it was difficult to properly assess differences in retrieval and control. Future research could use a larger set of closed questions to determine more precisely how retrieval, monitoring, and control contribute to the quantity and accuracy of eyewitness memory reports obtained via open-ended and closed questions.

DECLARATION

I certify that this thesis does not incorporate without acknowledgment any material previously submitted for a degree or diploma in any university; and that to the best of my knowledge and belief it does not contain any material previously published or written by another person except where due reference is made in the text.

Stacey Taylor-Aldridge

BBehavSci(Psych), BBehavSci(Hons)

ACKNOWLEDGMENTS

This thesis could not have been completed without the guidance and support of a great number of influential advisors, friends, family, and colleagues. Thankyou Nathan for the feedback and advice you have provided to me throughout this process and for helping me to recognise the value of my research, and to Neil for providing the lab space and equipment necessary to complete my experiments. To all the lab and office mates I've had over the years, thank you; Nicole for your organisation of the lab and interest in my project; Sarah and Nicky for giving me a place to vent, always believing in me, and helping me make sense of my data analysis, I truly do not think this would have been possible without you; Rosie for being my non-eyewitness office mate, occasional distraction, and great friend; Clare for your useful insights into my discussion sections in those last couple of months, despite the fact that you research a completely different area you always made a huge effort to understand and help; Fei for one of the best pieces of advice I have ever received: 'If nobody died, it can be fixed'; Carmen for being my travel companion for my first international conference; Sarah (again) for your advice and help with formatting and editing. Thank you also to Naomi, Rachel, Ashleigh, Stacey, Sophie, Sue, Jamie, and Amy for being my coffee/lunch buddies, providing me with much needed breaks, conversations about trashy TV, and a place to vent about teaching/marking frustrations. To my non-PhD friends (Aleks, Olwyn, Jacqui, Sarah and Ashley) and my friends from back home (Kirsty, Kaela and Abby) thank you for the much needed opportunities to relax with dinner, drinks, and/or chocolate, for making the effort to understand what the hell I've been doing for the past 4 years, and for being an excellent support network.

To Dan, my wonderful fiancé, I could not have asked for a better person to share my life with. While your rational thought sometimes gets in the way of my need to rant, I

appreciate that you try to find the best solution to a problem and always encourage me to hope for the best, even when it feels like everything is falling apart. Thank you also to my future in-laws, Lynne and Michael, for welcoming me into your family and always being willing to help Dan and I in any way you can. Finally, to my Dad, the words to describe my gratitude to you do not really exist. I know I have worked hard for what I have achieved but I never would have made it to this point without you. You saw potential in a child that most would probably have given up on and fought hard to make sure I had the opportunity to reach my potential. I thank you with all of my heart.

CHAPTER 1 – INTRODUCTION

Eyewitness interviews are an important part of the forensic investigative process and can provide investigators with invaluable information about a crime and its perpetrators. In fact, approximately 80% of police say that witnesses often or almost always provide the major leads in criminal investigations (Kebbell & Milne, 1998). Both the quality of eyewitness memory reports and the amount of information contained within them is important because police need to be able to trust that the information they obtain is correct and they need to acquire as much useful information as possible in order to conduct effective investigations. However, when people are given the opportunity to regulate the information they provide in their memory reports about an event they witnessed by choosing whether to volunteer or withhold particular details, they are unable to maximise both the quality and amount of information they volunteer (Weber & Brewer, 2008). In the metacognitive literature, the quality of a memory report is referred to as accuracy, and the amount of information obtained is referred to as quantity. Accuracy only considers information that is volunteered, and assesses how much of the volunteered information is correct (Koriat & Goldsmith, 1994). Quantity, however, considers all of the information that the witness could have encoded, and assesses how much of this information is both correct and volunteered (Koriat & Goldsmith, 1994). The overall aim of this thesis was to explore methods of improving the people's ability to simultaneously maximise the quantity and accuracy of their memory reports about a simulated crime. More specifically, the methods explored were designed to enhance metacognitive monitoring, which is the ability to discriminate between correct and incorrect memories.

Before I begin discussing the project in detail, a distinction must be made between the different test types and response formats that are commonly used to assess what people can remember about a witnessed event such as a crime. Three types of test can be used to

measure the information a witness can remember. The first type of test is recognition, in which witnesses are presented with an item that was either present or not present during the crime. For example, a shirt could be presented and witnesses are asked if it is the one worn by the offender. The second type of test that can be used is cued-recall, which provides witnesses with specific information to help guide their memory search. For example, witnesses may be asked what colour and type of shirt the offender wore. The third type of test that can be used is free-recall, which allows witnesses to guide their own retrieval with little input from the interviewer/experimenter. For example, witnesses may simply be asked to describe the offender's clothing. Recognition and cued-recall tests can be completed using two response formats; forced-report and free-report. Under forced-report conditions, witnesses are asked to answer all of the questions, while under free-report conditions they can choose to volunteer an answer for some questions and withhold an answer for others. Free-recall tests are generally only completed under free-report conditions because witnesses are not told what to retrieve. This means that is not possible to force them to volunteer every piece of information about a crime. However, they can be asked to volunteer everything that they do retrieve from memory as a variation of the forced-report procedure. In this thesis, my primary focus was on cued-recall tests completed under forced- and free-report conditions, with some exploration of free-recall tests in the final experiment.

In this thesis, I focussed on improving the way people respond to closed questions rather than open-ended questions. Closed questions can be answered with just a few words or, in some cases, a single word. For example, when asked 'What was the colour of the offender's shirt?' a witness could respond with the answer 'white'. Such questions form part of cued-recall tests. Open-ended questions, however, are used in free-recall tests and allow more extended responses. For example, when a witness is asked 'Can you describe

the offender?' their response could provide information about the offender's height, eye colour, clothing, and so on. Focussing on closed questions within this thesis was important from a practical research perspective because closed questions allow metacognitive monitoring to be assessed with ease. In general terms, assessments of metacognitive monitoring measure the association between accuracy (i.e., correct or incorrect) and confidence judgments and/or control decisions (i.e., decisions about whether to volunteer a piece of information or withhold it). Closed questions are designed to probe for information about specific details such as the colour of the shirt worn by an offender. Therefore, answers to closed questions can be easily coded as correct or incorrect because they typically relate to one discrete detail from the witnessed event. This also means that a confidence judgment can be requested following each question, and that the witness can be asked whether they want to volunteer or withhold an answer for each question. Therefore, it is easy to assess metacognitive monitoring when closed questions are used because confidence judgments and control decisions can be measured easily and are explicitly associated with a specific item in memory.

Although closed questions allow easy measurement of metacognitive monitoring and control, there are a several problems associated with their use that must be acknowledged. First, closed questions typically elicit less accurate memory reports than open-ended questions (Lipton, (1977). Fisher (1995) has proposed that this occurs because closed questions often assess aspects of the stimulus event that are more difficult to recall, a second problem associated with their use. When a closed question assesses a detail that a witness does not have a good memory for, Lipton (1997) has proposed that there is a stronger imperative to report the weak memory than if that same memory had been retrieved in response to an open-ended question. This is related to Fisher's (1995) argument that open-ended questions give witnesses more control over the information they

provide by allowing them to limit their responses to details for which they have a very good memory. This explanation also represents a third problem with closed questions; the limitations they place upon the information that can be offered by witnesses (Lipton, 1997). The use of closed questions means that witnesses have little opportunity to provide information that they are not specifically questioned about, which could result in important details going unreported. While these issues must be kept in mind when interpreting the findings presented in this thesis, and are discussed again in the final chapter, the use of closed questions was important for assessing the impact of various interventions on metacognitive monitoring and control.

Memory Regulation and the Quantity-Accuracy Trade-Off

Memory regulation in the context of eyewitness memory is the process by which people decide what information is volunteered in, and withheld from, their memory report about the crime. Koriat and Goldsmith's (1996) metamemory framework describes the memory regulation process in three stages; retrieval, monitoring, and control. When witnesses are asked a question, they begin the regulation process by attempting to retrieve information from memory. When they are successful in retrieving information, they monitor how likely the information is to be correct by making a subjective confidence judgment. They can then compare the confidence judgment to the level of confidence required to warrant volunteering an answer (i.e., a response criterion). Information is volunteered when confidence exceeds the response criterion and is withheld when it is below the response criterion.[1]

[1] This is not always the case as research by Ackerman and Goldsmith (2008) suggests that people sometimes violate their response criterion by choosing to volunteer information they do not feel sufficiently confident about. It was hypothesised that people do this to avoid appearing uninformative (Ackerman & Goldsmith, 2008).

The process of memory regulation can impact on the quantity and accuracy of people's memory reports. During the retrieval stage, quantity is reduced when a person cannot retrieve any information that answers the question.[2] The impact of monitoring judgments on quantity and accuracy depends on where the response criterion is set. When an incorrect piece of information is assigned a high confidence judgment, accuracy will be reduced when the response criterion is set at or below that confidence level. For example, if a person assigns an incorrect piece of information a confidence judgment of 90%, it will be volunteered if the response criterion is set at 90% or below, resulting in a reduced accuracy. Conversely, when a correct piece of information is assigned a low confidence judgment, quantity will be reduced when the response criterion is set above that confidence level. For example, if a person assigns a correct piece of information a confidence judgment of 50%, it will be withheld if the response criterion is set above 50%, resulting in reduced quantity. Even when all incorrect pieces of information are assigned low confidence judgments and all correct pieces of information are assigned high confidence judgments, control decisions can still impact on quantity and accuracy if the response criterion is not set appropriately. For example, imagine a person who assigns a confidence judgment of 80% or higher to correct information and a confidence judgment of below 80% to incorrect information. If this person sets their response criterion at 70%, accuracy will be reduced because some of the incorrect information they retrieved is assigned a confidence judgment above the response criterion. Conversely, if this person sets their response criterion at 90%, quantity will be reduced because some of the correct information they retrieved is assigned a confidence judgment below the response criterion.

[2] Failure to retrieve information only reduces quantity when the question relates to information that the witness could have encoded. If the question relates to information that was not present, or to details that the witness could not have encoded, failure to retrieve information will not reduce quantity. However, retrieving information about things that were not present or could not be seen can reduce accuracy.

Thus, failure to retrieve information as well as monitoring and/or control errors all contribute to the quantity and accuracy of people's memory reports.

It is important to improve monitoring so that errors in control, which contribute to reductions in quantity and accuracy, can be reduced. The need to optimise monitoring is illustrated when the outcome of monitoring and control errors is compared with ideal memory regulation. When performing optimally, people can maximise both the quantity and accuracy of their eyewitness memory reports. For example, consider a person who is asked 30 questions, all of which could be answered correctly (i.e., the questions do not ask about information that was not present; cf Scoboria, Memon, Trang, and Frey's (2014) work on unanswerable questions). They fail to retrieve an answer for four of these questions, retrieve a correct answer for 20 of them, and retrieve an incorrect answer for six of them. In order to maximise quantity and accuracy, this person would need to volunteer the 20 correct answers, withhold the six incorrect answers and say they do not know the answer to the four questions for which they could not retrieve an answer.[3] However, the person would only be able to do this if they can perfectly discriminate between correct and incorrect answers. More specifically, they can only maximise quantity and accuracy if there is no overlap between the confidence judgments assigned to correct and incorrect answers. This means, for example, that all correct answers are assigned a confidence level of 80% or above, and all incorrect answers are assigned a confidence level of below 80%. It is also essential that the response criterion is set appropriately, at 80% in this example. Metacognitive errors will arise when there is overlap in the confidence judgments assigned to correct and incorrect answers. If a person assigns 10 answers a confidence judgment of

[3] While this will not allow the witness to achieve 100% quantity as they did not retrieve a correct answer for all 30 questions, they would achieve the maximum quantity level, given their encoding and retrieval capabilities, of approximately 66% and an accuracy level of 100%.

80% and seven of these answers are correct, a response criterion of 90% will result in seven correct answers being withheld; reducing quantity but preserving accuracy. A response criterion of 70% will result in three incorrect answers being volunteered; reducing accuracy but preserving quantity. In these situations, control decisions are flawed due to monitoring errors and a quantity-accuracy trade-off occurs whereby increases in accuracy come at the cost of reduced quantity and vice versa. Therefore, in this thesis, I focus on monitoring errors because they precede and contribute to control errors which can ultimately affect quantity and accuracy.

There are two different ways of assessing monitoring which can only be optimised simultaneously in a particular set of circumstances. The first type of monitoring is calibration (Adams, Smith, Pasupathi, & Vitolo, 2002); which is defined as the degree of match between confidence judgments and actual accuracy (see, e.g., Williamson, Weber, & Timmins, 2012). When a person is perfectly calibrated, 100% of the details to which they assign 100% confidence are correct, 90% of the details to which they assign 90% confidence are correct, and so on. The second type of monitoring is resolution; which is broadly defined as the ability to discriminate between correct and incorrect information (see, e.g., Williamson et al., 2012). To achieve perfect resolution, each confidence judgment must be assigned uniquely to either correct details or incorrect details. For example, perfect resolution can be achieved when all correct details are assigned a confidence level of 80% or above and all incorrect details are assigned a confidence level of below 80%. In this example, although resolution is perfect, calibration is not because all details that are assigned an 80% confidence judgment are correct, rather than 80% of the details assigned to the 80% level being correct. Similarly, when calibration is perfect, resolution is usually not because confidence judgments are not assigned uniquely to either correct or incorrect details. The only circumstance in which both perfect calibration and

perfect resolution can be achieved is when all correct details are assigned 100% confidence and all incorrect details are assigned 0% confidence (Williamson et al., 2012).

In order to enable people to make good control decisions, optimising resolution is vital because even when calibration is perfect, it may not be possible to maximise quantity and accuracy. As an example, consider a person with perfect calibration who sets a response criterion of 80%. As they are perfectly calibrated, 20% of the details to which they assign 80% confidence and 10% of the details to which they assign 90% confidence will be incorrect. This incorrect information will be volunteered because their response criterion is set at 80%, causing a reduction in accuracy. Quantity will also be reduced in this example because all correct answers assigned a confidence level of below 80% will be withheld. However, if resolution was perfect, it would be possible to maximise quantity and accuracy. If the person in the example assigned all correct details a confidence judgment of 80% or above and all incorrect details a confidence judgment of below 80%, they would maximise quantity and accuracy because their response criterion is set at 80%. They are able to achieve maximum quantity and accuracy despite the fact that their calibration is imperfect. Thus, perfect resolution is generally more useful than perfect calibration as a basis for control decisions.[4] For this reason, I chose to focus on monitoring

[4] This does not mean that calibration is entirely unimportant in the context of control decisions. For example, a person can achieve perfect resolution by assigning all correct details a confidence level of 60% or above and all incorrect details a confidence level of below 60%. In this situation, confidence does not realistically represent the likelihood that the answer is correct. Nevertheless, accuracy and quantity will be maximised when the person sets their volunteering criterion at 60%. However, some degree of calibration is required so that the criterion can be set appropriately. In this situation, the person needs to be aware that they are underconfident in many cases (e.g., that they have assigned some correct details a confidence level of just 60%).

resolution in this thesis as opposed to calibration, with the ultimate aim of helping people maximise the quantity and accuracy of their eyewitness memory reports.

The focus on monitoring resolution is also important because imperfect monitoring resolution has been observed in both general knowledge tests and eyewitness recall tasks. Koriat and Goldsmith (1996, Experiment 1) had participants answer a series of general knowledge questions to assess monitoring resolution. They observed a mean adjusted normalised regression index (ANDI)[5] of .61 and a mean Goodman-Kruskal gamma correlation[6] (*G*) of .87, indicating good but imperfect monitoring. Imperfect monitoring was also observed by Weber and Brewer (2008, Experiment 2) in an eyewitness memory task in which participants were presented with questions about a video they had viewed and provided a fine-grained answer (i.e., a specific answer such as the exact colour of the shirt worn by an offender) and a course-grained answer (i.e., a general answer such as the general tone of the shirt worn by an offender) for each. In a second phase of the experiment, participants were given the option of deciding whether they wanted to volunteer one of these answers or withhold both of them. Weber and Brewer (2008) observed that confidence discriminated between volunteered and withheld answers in a similar way for fine-grain (mean ANDI = .32; mean *G* = .74) and coarse-grain (mean ANDI = .24; mean *G* = .70) responses, but that monitoring ability was moderate to low. Furthermore, when Perfect (2004, Experiment 2) directly compared monitoring during an eyewitness recall task and a general knowledge test, he observed a higher gamma

[5] ANDI scores can range from 0 (indicating no discrimination between correct and incorrect answers) to 1 (perfect discrimination between correct and incorrect answers).

[6] Goodman-Kruskal gamma correlations can range from -1 (indicating that all incorrect answers are assigned a higher confidence rating than all correct answers) to +1 (indicating that all correct answers are assigned a higher confidence rating than all incorrect answers).

correlation for general knowledge (G = 0.74) than for eyewitness memory (G = 0.58). As these findings indicate that people are unable to perfectly monitor the accuracy of their memories, particularly in eyewitness memory tasks, there is scope for improving monitoring ability.

There is also evidence which suggests that monitoring can be manipulated, and that doing so can affect quantity and accuracy. Koriat and Goldsmith (1996, Experiment 2) manipulated the monitoring ability of participants by presenting them with a general knowledge test that contained deceptive items (i.e., questions that produce an illusion of knowing) in addition to standard items. In phase one of the experiment, participants were asked to answer all of the questions (i.e., forced-report), while in a second phase they could choose which questions to answer (i.e., free-report). Koriat and Goldsmith's (1996) results showed that the monitoring manipulation was successful given that poor monitoring was observed for the deceptive items (mean ANDI = .03; mean G = .26) and good monitoring was observed for standard items (mean ANDI = .64; mean G = .90). This difference in monitoring had important implications for accuracy. When monitoring was good (i.e., standard items), participants were able to increase their accuracy by a mean of 47.1%. However, when monitoring was poor (i.e., deceptive items), participants were only able to increase their accuracy by a mean of 9.8%. Thus, when monitoring was poor, participants were less able to improve their accuracy. While quantity was reduced by a similar percentage regardless of monitoring ability (by a mean of 4.2% and 5.6% when monitoring was poor and good, respectively), the fact that there was any drop in quantity illustrates that imperfect monitoring can have a detrimental impact on quantity. Although these results are from a test of semantic memory, it is likely that monitoring would have a similar, if not more severe impact on quantity and accuracy during an episodic memory test given that monitoring tends to be less effective in eyewitness tasks compared to

general knowledge tasks (Perfect, 2004). Therefore, exploring ways of improving monitoring ability represents an important opportunity to help people maximise the quantity and accuracy of their eyewitness memory reports.

Improving Monitoring: The Effects of Feedback, Warnings, and Training

Although relatively little research has been conducted into techniques that can improve monitoring, several studies have examined the effect of performance feedback. Performance feedback involves providing participants with information about the accuracy of their judgments (Benson & Önkal, 1992; Sharp, Cutler, & Penrod, 1988; Stone & Opel, 2000). The provision of performance feedback has produced mixed results. Stone and Opel (2000) tested the effect of performance feedback (i.e., a calibration graph and suggestions for improvement) on participants' ability to decide whether a piece of artwork was from the later of two presented time periods. This feedback did not improve participants' resolution, though it did improve calibration. It is likely that the feedback failed to improve resolution because it was related to calibration. However, Benson and Önkal (1992) also failed to improve resolution using performance feedback that was aimed specifically at resolution. In their study, participants were asked to make predictions about the outcome of upcoming football matches and indicate the likely accuracy of these predictions. The study took place over four sessions with feedback on the previous session given at the beginning of each session. Participants were given one of three types of feedback; calibration (i.e., calibration score and curve), resolution (i.e., calibration curve and resolution score), and covariance (i.e., Brier score[7] and covariance graph). All participants received outcome feedback (i.e., their rank within the group based on Brier scores), and a fourth group only received outcome feedback. Feedback did not improve resolution in any of the four conditions. Conversely, Sharp et al. (1988) did find an

[7] The Brier score can be considered a global monitoring measure that reflects both resolution and calibration.

improvement in resolution following performance feedback. Participants in their study completed a general knowledge test and were then provided with information about the relationship between their confidence in the answers and their actual accuracy. Specifically, participants were given a table containing the proportion of correct answers assigned to each confidence judgment, along with the number of judgments they made for each confidence level, and the average probability of making a correct judgment. Findings showed that resolution improved when participants were given this feedback.

Despite the promising results of Sharp et al. (1988), the provision of performance feedback was not a viable option in this thesis because it is not practically feasible in the applied eyewitness context (Lane, Roussel, Villa, & Morita, 2007). This was an important factor to consider because my aim was to develop techniques that could be used in more naturalistic eyewitness settings in the future. By definition, performance feedback requires prior knowledge of the information/context in question so that training can be provided to improve performance on the task of interest. While such knowledge is available in laboratory based studies such as those conducted in this thesis, the information being requested in natural eyewitness settings is unknown to the interviewer (Lane et al., 2007). Thus, while feedback could be given in laboratory studies during a training phase, it could not be given in real police interviews because the accuracy of the information in question is unknown. Furthermore, Bornstein and Zickafoose (1999) have also found that giving performance feedback on a general knowledge test does not improve resolution on a subsequent eyewitness memory test. This means that any positive effects of feedback do not appear to generalise, and providing feedback on a contrived task (i.e., a general knowledge task) does not improve resolution on the real task of interest (i.e., an eyewitness task). Thus, because providing feedback is not viable in naturalistic eyewitness settings

and the effects of feedback do not appear to generalise across tasks, this thesis did not explore feedback as a method for improving monitoring.

Although warning people about possible memory errors has sometimes been found to reduce false recognition, inconsistent results mean that such warnings may not be a useful means of improving monitoring. The Deese-Roediger-McDermott (DRM) paradigm involves presenting participants with a list of words that are all associated with a critical theme word. For example, participants may be sequentially presented with the words *smile*, *laugh*, and *fun* which are associated with the critical theme *happy*. When subsequently presented with an old-new recognition test, participants often falsely recognise the critical theme word as old (Roediger & McDermott, 1995). This is known as the DRM false recognition effect (Gallo, Roediger, & McDermott, 2001). While two studies have found that the DRM false recognition effect can be reduced by warning participants about the nature of the lists after they have been studied (McCabe & Smith, 2002; Starns, Lane, Alonzo, & Roussel, 2007), the majority of the research on such post-study warnings suggests that they do not significantly reduce the DRM false recognition effect (Anastasi, Rhodes, & Burns, 2000; Gallo et al., 2001; Miller, Guerin, & Wolford, 2011; Neuschatz, Payne, Lampinen, & Toglia, 2001).[8] However, Anastasi et al. (2000) and Neuschatz et al. (2001) did find that post-warnings lower the level of confidence participants have when they falsely recognise critical theme words. Thus, it appears that while post-study warnings can alter the way people experience falsely recognised words,

[8] Although numerous studies have found that warning participants about the nature of the lists before they are studied can reduce the DRM false recognition effect(Gallo, Roberts, & Seamon, 1997; Gallo, Roediger, & McDermott, 2001; Jou & Foreman, 2007; McCabe & Smith, 2002; McDermott & Roediger, 1998; Multhaup & Conner, 2002; Neuschatz, Benoit, & Payne, 2003; Roediger & McDermott, 1995; Watson, McDermott, & Balota, 2004), such pre-study warnings are not useful in natural eyewitness settings because a warning about memory errors cannot be given before a crime occurs.

they do not consistently produce better performance on the DRM task. This suggests that basic warnings may not improve monitoring.

This conclusion is also supported by the inconsistent results that have been observed when warnings are used to help witnesses avoid volunteering false memories that were suggested to them after witnessing the event. Studies of suggested memories involve showing participants an event and subsequently suggesting false information to them about that event (i.e., misinformation). After receiving the misinformation, participants are asked to complete a recognition, cued-recall, or source memory test. Source memory tests ask participants to distinguish between items they actually saw, items that were only suggested to them, and items that were both suggested to them and actually seen. Participants in these studies frequently indicate that the misinformation was part of the event they witnessed (Loftus, 2005). This is termed the misinformation effect (Tousignant, Hall, & Loftus, 1986). Some research suggests that the misinformation effect can be reduced (and sometimes eliminated) by providing participants with a warning about the presence of inaccurate information within the post-event information after it has been presented (Christiaansen & Ochalek, 1983; Echterhoff, Groll, & Hirst, 2007; Oeberst & Blank, 2012; Szpitalak & Polczyk, 2010). However, other research has found that such post-misinformation warnings do not have a significant impact on the frequency with which witnesses volunteer misinformation (Greene, Flynn, & Loftus, 1982; Higham, Luna, & Bloomfield, 2011; Zaragoza & Lane, 1994).[9] Furthermore, Echterhoff et al. (2007) and

[9] Higham et al. (2011) concluded that the post-misinformation warning was ineffective because the misinformation effect was not eliminated (i.e., participants continued to volunteer some misinformation after they received the warning). However, their experiment did not include a condition that did not receive a post-misinformation warning. Thus, it was not possible to determine whether the warning significantly reduced the misinformation effect.

Szpitalak and Polczyk (2010) have found that witnesses often withhold correct information about the event that was presented in the post-event information in addition to withholding the misinformation. Thus, post-misinformation warnings may not always reduce the misinformation effect and they may reduce the amount of correct information that is volunteered. Therefore, on the basis of these results and those observed in the DRM studies reported above, this thesis did not initially explore warnings about the potential inaccuracies of memory as a way of improving monitoring, though Experiment 4 did address this possibility in light of the results from Experiments 2 and 3.

Although feedback and warnings did not appear to be viable options for improving eyewitness monitoring, prior research also suggested that training aimed at improving knowledge of the task might be useful. For example, Lichtenstein and Fischhoff (1977) gave their participants five minutes of training designed to increase knowledge about the difference in handwriting between American and European adults. Specifically, the training allowed participants to study examples of correctly labelled stimuli, and as a result of this small amount of training, resolution, calibration and overall accuracy for the task were improved. Similarly, Stone and Opel (2000) also used training to improve resolution, although they termed their training *environmental feedback* (i.e., providing information about the task upon which judgments are based). Participants received a 30-minute lecture about art history which outlined the important characteristics of art from four different periods. Following this, participants were presented with slides of artwork and asked to judge if they were from the later of two periods presented on the slide. The training improved resolution in comparison to the pre-training testing. Together, Lichtenstein and Fischhoff (1977) and Stone and Opel's (2000) findings illustrate that increasing people's knowledge of the content of a task can improve their monitoring.

Although increasing people's knowledge of the content of a task through training can improve monitoring, such a technique was inappropriate for this thesis because it could not be used in real eyewitness interviews. As explained earlier, this was an important consideration given my aim of developing techniques that could be used in more naturalistic eyewitness settings in the future. In the naturalistic eyewitness setting, it is difficult to give people more information about the content of the task because the content is the witnessed event. If the content was known, there would be no need to interview the witnesses at all. However, it may be possible to train people to become more knowledgeable about factors that discriminate between correct and incorrect memories, rather than training them about the task content (Lane, Roussel, Starns, Villa, & Alonzo, 2008; Lane et al., 2007; McCabe & Soderstrom, 2011). Such a technique, if successful, has the potential to be applied in real eyewitness interviews in future research.

Improving Monitoring: The Use of Mnemonic Cues

Telling people what information they should use to distinguish between correct and incorrect memories is a potentially useful method for improving monitoring because people may not spontaneously use information that is most diagnostic of accuracy. McCabe and Soderstrom (2011) found that when people are asked to make prospective memory judgments (i.e., determine whether they will remember studied information), their judgments are more accurate when they are told to base them on a particular kind of information (i.e., the presence or absence of contextual details) than when they are allowed to freely determine what information to consider while making their confidence judgments. This finding illustrates that people do not always automatically base their judgments on information that is diagnostic of later retrieval. However, it also illustrates that, at least in some situations, people do have access to more diagnostic information. Therefore, when

assessing the accuracy of a memory, people may need to be informed about what information they should attend to because they do not make use of it spontaneously.

The source monitoring framework (Johnson, Hashtroudi, & Lindsay, 1993) describes the different sources memories can come from and outlines the types of available information (i.e., mnemonic cues) that help distinguish between memories from different sources. According to the framework, memories can be created by both external stimuli and thought processes, and thus can have an external or internal origin, respectively (Johnson et al., 1993; Johnson & Raye, 1981). For eyewitnesses, an external memory is an element of the crime scene (e.g., the getaway car), and I proposed that correct memories about a witnessed event may be analogous to the external memories described in the framework because they are created by an external source. Conversely, I proposed that incorrect memories may be analogous to the internal memories described in the framework because they are likely to be the result of internal cognitive processes (e.g., a witness may imagine that the offender used a bag to carry stolen money because this is consistent with the witness's schema of what occurs during a robbery).[10] A variety of mnemonic cues can help people distinguish between internally and externally generated memories including: (i) sensory information (e.g., visual and auditory), (ii) contextual detail (e.g., spatial and

[10] Of course, it is also possible for incorrect memories to be created by an external source. For example, a person could be exposed to an incorrect piece of information about the crime they witnessed via a news report. If this incorrect piece of information comes to mind during a subsequent police interview, the witness must determine that it appeared in the news report, and not during the actual event, if they are to avoid reporting it. These distinctions differ from monitoring decisions which relate to accuracy, and, as such, are beyond the scope of this thesis. Nonetheless, it is important to keep in mind that in some circumstances judgments about the likely accuracy of a memory will also require the witness to make a decision about where that memory came from.

temporal), (iii) semantic content, (iv) affective information, and (v) cognitive processes (Johnson et al., 1993; Johnson & Raye, 1981). Cognitive processes can include thoughts, associations, imagination, reasoning, decisions, and elaboration (Johnson et al., 1993; Johnson, Kahan, & Raye, 1984; Johnson & Raye, 1981), though no clear definition of cognitive processes is given in the literature. These mnemonic cues and their associations with internal and external memories can help people to determine the origin of a particular memory (Johnson & Raye, 1981).

The processes people can use to discriminate between memories from internal and external sources are outlined in the source monitoring framework. When determining the source of their memories, people will often make quick and automatic decisions that are based on a superficial assessment of the mnemonic cues associated with those memories (Johnson et al., 1993). However, in addition to these heuristic source judgments, people are also capable of engaging in more strategic processes in which they consider the mnemonic cues more carefully (Johnson et al., 1993). It is these systematic source judgments that can alert people to inaccuracies in their heuristic source judgments (Johnson et al., 1993). Therefore, as systematic source judgments can assist people in accurately differentiating between externally and internally generated memories, and I proposed that external and internal memories are analogous to correct and incorrect memories, respectively, encouraging systematic assessments of mnemonic cues may improve discrimination between correct and incorrect memories.

Research on autobiographical memory has identified mnemonic cues that can distinguish between perceived and imagined autobiographical events. Johnson, Foley, Suengas, and Raye (1988) asked participants to remember both a perceived event and an imagined event and complete a memory characteristics questionnaire that assessed a wide range of mnemonic cues (e.g., sensory detail, complexity, spatial, temporal, and contextual

information, and feelings). The findings showed that in comparison to imagined events, perceived events contained more sensory details (i.e., visual, auditory, and olfactory) and more spatial, temporal, and contextual information. In a second study, Johnson et al. (1988) also asked participants to remember a perceived and an imagined event. Rather than completing a memory characteristics questionnaire they were asked to describe how they knew the perceived memory had actually occurred and how they knew the imagined event had not. For perceived events, participants indicated that they knew these events had occurred because they could recall characteristics of the memory and supporting memories which were related to the target event. However, for imagined events, participants indicated that they knew these had not occurred because the memory involved reasoning. Together, these findings suggest that memories for events that did occur can be identified by sensory and contextual information while memories created through imagination can be identified by cognitive processes.

There are also differences in mnemonic cues associated with memories for events that did occur and memories created through misinformation that can help people discriminate between them. As explained earlier, studies of the misinformation effect involve presenting participants with an event, giving them misinformation about that event, and asking them to compete a memory test. Schooler, Gerhard and Loftus (1986) showed participants a series of slides depicting a traffic accident and manipulated whether or not a yield sign was present. Post-event misinformation told participants who did not see the sign that they had seen it. In a source memory test, participants were asked to provide a description of their memory for the sign when they indicated that it had been seen in the slides. The findings showed that those who did not see the sign were more likely to describe cognitive processes, describe the function of the sign, and use verbal hedges (e.g., I think) than participants who actually saw the sign who reported more sensory details (i.e.,

colour, size, shape). Schooler et al. (1986) also found that providing people with a summary of the differences between memories for an event and suggested memories improved their ability to distinguish between these two types of memory in other people. Furthermore, Lane et al. (2007) found that informing participants of the differences between memories for an event and suggested memories (i.e., clear memory for the item and its location) improved their ability to distinguish between their own real and suggested memories. In addition, Bulevich and Thomas (2012) found that encouraging participants to consider visual imagery and auditory and contextual information was able to improve resolution and accuracy on a recognition task following the provision of misinformation. Together, these results show that memories for events that did occur are associated with sensory information while suggested memories are associated with cognitive processes. Moreover, knowledge of these differences can help people discriminate between these two types of memory in others and within themselves.

Similarly, studies of the DRM paradigm have identified differences in the mnemonic cues associated with studied words and critical theme words that can help people make this distinction for themselves. As explained earlier, the DRM paradigm involves presenting participants with a list of words which are all associated with a critical theme word. On a subsequent recognition test, participants often falsely recognise the critical theme word as old. Norman and Schacter (1997) asked participants to provide explanations for their recognition judgments to explore the mnemonic cues associated with critical theme words and studied words. They found that explanations regarding critical theme words referred to related words and thoughts or associations to the word while explanations of the studied words referred to the word's presentation (i.e., reactions, context and sensory characteristics). In addition, Lane et al. (2008) found that when explicitly informed of the mnemonic cues associated with old items in the DRM paradigm

(e.g., good memory for how the word sounded and where it was located in the list), participants' performance on the task improved, although Neuschatz, Benoit, and Payne (2003) failed to find such an effect. Together, these results show that studied words are associated with sensory and contextual information while theme words are associated with cognitive processes. Furthermore, informing people of these differences can sometimes allow them to perform better on the DRM task because they are better able to discriminate between words they studied and words that are closely associated with words they studied.

Overall, research has found that memories of presented words/images and memories for event that did occur are associated with sensory and contextual information, while cognitive processes (e.g., thoughts, associations, reasoning, and elaboration) are associated with memories created through misinformation, imagination, or association. Importantly, these associations are more than just descriptive: Informing people of these mnemonic cues can help them perform better on recognition tasks in the DRM and misinformation paradigms. Thus, there is evidence that useful mnemonic cues exist and that they can improve people's performance in basic memory tasks. Therefore, it may be possible to improve peoples' monitoring ability in a complex episodic memory task by informing them of differences in the mnemonic cues associated with correct and incorrect memories.

Overview

The main purpose of the studies presented in this thesis was to develop a technique that would help people simultaneously maximise the quantity and accuracy of their eyewitness memory reports when closed questions are asked. I aimed to achieve this by improving monitoring ability (i.e., resolution). Research suggests that it may be possible to do this by increasing knowledge of the different mnemonic cues that are associated with correct and incorrect memories. To assess whether this could be achieved, Study 1 aimed

to establish that the mnemonic cues identified in previous research can discriminate between naturally occurring correct and incorrect memories (i.e., memories not created via suggestion or deliberate imagination) in a complex eyewitness memory task (Chapter 3). In Experiments 2 and 3, knowledge of the mnemonic cues was manipulated and the impact of these manipulations on monitoring, quantity, and accuracy was assessed (Chapter 4). As these experiments suggested that witnesses may overestimate the accuracy of their memory, which could result in lax monitoring, Experiment 4 manipulated knowledge of a mnemonic cue which appeared to be the most reliable predictor of accuracy across the other three studies, in addition to warning participants of the fallibility of eyewitness memory (Chapter 5). However, as the results of Experiments 2 and 3 could have also indicated that closed questions discourage withholding, Experiment 5 explored whether retrieval, monitoring, and control abilities differ depending on whether open-ended or closed questions are used (Chapter 6).

CHAPTER 2 – ANALYTICAL APPROACHES

This chapter describes the mixed effects modelling approach that was used to analyse the data obtained from five studies, the results of which are presented in chapters 3-6. In this chapter I will: (i) explain the value of this approach for analysing the type of data yielded in the studies that were conducted which contained multiple items and, in one instance, multiple memory tests; (ii) outline the problems with the approaches traditionally used to analyse this type of data, and explain how they are addressed by the analytical approach used throughout this thesis.

Limitations of Traditional Approaches for Studies with Multiple Items

Due to the nature of the data obtained in the studies presented in this thesis, traditional approaches to analysis were inappropriate. In each of my studies, participants saw the same video of a mock crime and responded to multiple questions about their memory for this stimulus, each coded as either correct or incorrect. Depending on the study, participants provided ratings for the answers they gave (e.g., confidence ratings and mnemonic cue ratings) and/or decided whether each answer would be volunteered or withheld. Traditional approaches to analysing data of this type use aggregate statistics such as mean confidence, mean proportion correct, and mean *d*'.[11] These aggregate statistics are analysed using *t*-tests or ANOVAs to assess differences between groups. For proportions, this is particularly problematic because the assumption of a normal distribution is violated and equal variance cannot be assumed because the range of scores lies between 0 and 1 (Dixon, 2008; Quené & van den Bergh, 2008). Furthermore, ceiling and floor effects are common because of the restricted range of scores (Dixon, 2008).

[11] *d*' is commonly used to assess recognition performance. In this thesis, I used a Type-2 Signal Detection Theory approach which is described in detail on p. 36.

Similarly, some aggregate statistics (i.e., *d*') cannot be calculated when performance is at the extremes, such as when a participant has a hit rate or false alarm rate of 0% or 100%. In instances where this occurs, adjustments must be applied to the data meaning the aggregate does not fully represent participant responses (Higham & Tam, 2005), or the data needs to be excluded from the analyses entirely (Murayama, Sakaki, Yan, & Smith, 2014). Aggregate statistics are also problematic because they can mask effects that are present or make effects that are not present appear as though they are (Baayen, Davidson, & Bates, 2008). This is partially because the traditional approaches to analysing aggregate statistics fail to account for variation in the effect of manipulations between items and between participants at the same time.

Traditional analyses are also problematic for Type 1 errors and generalisability. The questions included in the studies presented in this thesis did not exhaust all of the possible questions that could have been asked, and particular questions may have been more difficult to answer than others. Furthermore, all of the questions were linked to the particular video that was used, which may differ from questions that could be asked about other stimulus events. Significant variation could also exist for the relationships under investigation between questions. For example, a particular mnemonic cue could be a strong predictor of accuracy for some questions but a weak predictor for others, while a different mnemonic cue may be a consistently moderate predictor for all questions. Any analysis that treats these mnemonic cues in the same way, without accounting for variability, could produce statistical errors and lead to theoretical and applied misinterpretations. Specifically, because traditional analyses are not capable of accounting for differences in the effects being investigated across items, they can inflate Type 1 error rates (Murayama et al., 2014). There will also be issues with generalising the results of a particular study because the *p*-values produced by traditional analyses are only relevant for

the items included within that study (Westfall, Kenny, & Judd, 2014). This is obviously problematic from an applied perspective because experimental results are only useful if they can be generalised to the real world. For example, knowing that a particular mnemonic cue predicts response accuracy for the exact questions asked about one particular stimulus event viewed under a specific set of conditions is only helpful from an applied perspective if this mnemonic cue also predicts accuracy for other questions about other stimuli viewed under various conditions. The only type of analyses that are technically able to support such a generalisation are those that account for variation between items, such as mixed effects approaches (Westfall et al., 2014).

Mixed Effects Modelling

An appropriate alternative to the traditional methods of analysis is mixed effects modelling. This analysis approach does not use aggregate statistics, and as a result, provides a better representation of all the data. This approach addresses the limitations associated with the use of aggregate statistics because each data point (e.g., each answer) for every participant makes a contribution to the model. All data can be used from all participants, and any effects of differences between items and participants can be accounted for in the model. The resulting analyses are also more powerful than the traditional approaches outlined above (Baayen et al., 2008; Jaeger, 2008; Judd, Westfall, & Kenny, 2012; Kliegl, Wei, Dambacher, Yan, & Zhou, 2011; Quené & van den Bergh, 2008) and the risk of Type 1 error is reduced (Barr, Levy, Scheepers, & Tily, 2013; Garson, 2012; Jaeger, 2008; Judd et al., 2012; Murayama et al., 2014; Quené & van den Bergh, 2008). The following sections provide a more detailed explanation of mixed effects modelling, and describe how it is used throughout this thesis.

Mixed effects modelling is an extension of regression with a model created to estimate the change in an outcome measure with changes in a predictor variable/s (Baayen

et al., 2008; Jaeger, 2008; Kliegl et al., 2011; Murayama et al., 2014; Quené & van den Bergh, 2008; Winter, 2013). However, while normal regression has just one error term that is assumed to account for all variation not caused by the predictor, mixed effects models allow part of the error variance to be explicitly modelled as being the result of variation within specified variables (e.g., in this thesis, variation between questions or participants). The name mixed effects modelling reflects the fact that these models include both fixed and random effects (Bates, 2007). Fixed effects are factors for which every variation in the population is represented in the study (e.g., experimental manipulations). Random effects are factors for which only a sample of the entire population is used (e.g., specific participants and particular questions about a specific stimulus). Variation between participants and items that does not reflect the effect of the variables of interest can be accounted for by adding these factors to the model as random effects. This is done by allowing regression parameters (i.e., the intercept or the intercept and slope) to vary randomly for participants and items. When a variable is entered as a random intercept, the model estimates the variability in intercept by effectively estimating a different intercept coefficient (which indicates the base level of performance, ignoring the impact of other predictors) for each level of the random effect.[12] Similarly, when a variable is entered as a random slope, the model estimates the variability in slope by effectively calculating a different slope coefficient (which indicates the amount of change in the outcome measure with one unit change in a predictor) for each level of the random effect.

[12] Non-Bayesian mixed effects modelling does not actually estimate the parameter for each level of the random effect in the same way that the overall regression coefficient is estimated. Technically, only the variance in the regression coefficient is considered as a statistical parameter in the model. However, this simplification provides a sensible conceptual-level description of the operation of these models, even though it misrepresents some of the underlying mathematical complexity.

In all the models constructed in this thesis, question was entered as a random intercept and random slope because each question occurred in every level of the fixed effects. In other words, all participants received the same set of questions regardless of the experimental condition they were assigned to.[13] Participant was also included as a random intercept and slope in all within-subjects analyses. The inclusion of random slopes for participant is appropriate for predictor variables that are measured or manipulated within-subjects because the effect of the predictor may vary across participants. For predictor variables that are manipulated between-subjects, each level of the manipulation is completed by a different group of participants. Thus, differences between the participants reflect differences between the levels of the manipulation, rather than differences in how participants respond to different levels of the manipulation. For this reason, only a random intercept for participant was included in between-subject analyses.[14]

The mixed effects models reported in this thesis were created using the lme4 package (Bates, Maechler, Bolker, & Walker, 2015) in R, an open source language and environment for statistical computing (R Development Core Team, 2015). The outcome in all models was either accuracy (binomial, where 1 is correct and 0 is incorrect), quantity (binomial, where 1 is volunteered a correct answer and 0 is did not volunteer a correct answer) or control decision (binomial, where 1 is volunteered and 0 is withheld). As all

[13] If participants had received different questions based on the experimental condition they were assigned to, including a random slope for question would be inappropriate because the effect of condition could not vary based on question, as each condition would receive only one set of questions.

[14] The models constructed in Study 1 only included random intercepts for participant and question as the importance of including random slopes has only been convincingly demonstrated in recent literature (e.g., Murayama et al., 2014; Westfall et al., 2014) and was, therefore, not apparent at the time that these analyses were run. As the design of subsequent studies was based largely upon the Study 1 results, I did not re-run these analyses.

outcome measures were dichotomous, a logit link function was most appropriate. Thus, the mixed effects modelling used in this thesis is an extension of logistic regression rather than normal regression. This means that the probability of the outcome given a certain value of the predictor is estimated (Jaeger, 2008).

When constructing each model, the random effects were entered first, followed by the fixed-effect predictor/s.[15] When multiple predictors were entered, they were added in order from those of least interest to those of most interest. For each predictor added to the model, the models generated a regression coefficient (*b*) and a standard error for the regression coefficient (SE_b). In logistic regression, the regression coefficient represents the amount of change in the log odds of the outcome with one unit change in the predictor when all other variables are zero. When categorical predictors with more than two levels (e.g., experimental condition) are entered, the categories are considered in alphabetical/numerical order with the first category used as the reference category to which all other categories are compared (Winter, 2013). For example, when there are three conditions such as control, experimental condition A and experimental condition B, the reference condition will be the control condition, and experimental condition A and experimental condition B will be the comparison conditions. While the model takes all of the conditions into account when the parameters are estimated, it does not provide an explicit significance test of the difference between the comparison conditions.[16] Thus, the analysis that uses the control condition as the reference will produce two regression

[15] All continuous variables were centred prior to analysis. Throughout the thesis, I use the term *predictor* to refer to fixed-effect predictors in contrast to *random effects*.

[16] In order to obtain a significance test for the difference between the comparison conditions, a separate analysis needs to be conducted which includes only the comparison conditions (e.g., experimental condition A and experimental condition B). In this case, experimental condition A will be used as the reference condition and experimental condition B will be the comparison condition.

coefficients. One coefficient will represent the change in the slope of the line of best fit between the control condition and experimental condition A. The other coefficient will represent the change in the slope of the line of best fit between the control condition and experimental condition B.

I assessed significance for the mixed effects models in two ways. The first involved assessing whether model fit was significantly improved with the addition of a particular predictor variable. Specifically, a model containing the predictor of interest was compared to a model without that predictor (Jaeger, 2008; Winter, 2013). For example, when assessing the effect of different instructions on accuracy, a model containing instruction type as the predictor (in addition to a random intercept for participant and a random slope and intercept for question) is compared to a model containing only the random effects. A significant difference between the two models indicates that the more complex model is a better fit to the data (Jaeger, 2008; Winter, 2013). All models presented in this thesis were compared to a baseline model that included a random intercept and slope for participant and random intercept and random slope for question. The only exceptions to this were analyses that included a predictor variable that was manipulated within-subjects. In these analyses, only a random intercept was included for participant, along with a random intercept and slope for question. The second way I assessed significance was with 95% confidence intervals (CI's) to determine whether the *b* values were significantly different from zero. Confidence intervals were calculated by multiplying the standard error of the coefficient (SE_b) by 1.96 and adding (upper limit) this value to, or subtracting (lower limit) this value from the coefficient (*b*) (Gelman & Hill, 2006). When the 95% CI for an effect did not include zero (alpha level $\leq .05$), it was interpreted as significant.

To further explain the mixed effect modelling approach, consider an example from Chapter 4. Table 1 shows a set of model coefficients from Experiment 2. This experiment considered whether instructions about mnemonic cues could improve monitoring and whether such instructions impact on the quantity and accuracy of eyewitness memory reports. The results presented are from an analysis that compared accuracy between a condition that received instructions about confidence and three conditions that received instructions about different mnemonic cues. While this analysis took all of the conditions into account when the parameters were estimated, it did not provide an explicit significance test of the difference between the comparison conditions, as explained earlier. Thus, for each of the mnemonic cue conditions, the table lists the regression coefficient (*b*), the standard error for that coefficient (SE_b) and the 95% confidence interval for the coefficient ($95\%CI_b$). The *b* values indicate the extent to which each mnemonic cue condition differed from the confidence condition in terms of accuracy. A positive value indicates that the mnemonic cue condition obtained higher accuracy than the confidence condition, while a negative value indicates that the mnemonic cue condition obtained

Table 1

Fixed Effects Coefficients for Accuracy for the Clarity + Reasoning, Clarity + Thoughts, and Visual Detail + Effortfulness Conditions in Comparison to the Confidence Condition in Experiment 2

Condition	*b*	SE_b	95% CI_b
Clarity + reasoning	0.10	0.31	-0.50, 0.70
Clarity + thoughts	-0.14	0.25	-0.63, 0.35
Visual detail + effortfulness	-0.19	0.26	-0.69, 0.31

lower accuracy than the confidence condition. Thus, in comparison to the confidence condition, accuracy was higher in the clarity + reasoning condition but lower in the clarity + thoughts and visual detail + effortfulness conditions. However, as the 95% CIs include zero, none of the differences were significant. Consistent with this, a model containing condition as a predictor variable did not fit the data significantly better than a model that did not contain condition as a predictor variable, χ^2 (3) = 0.88, p = .830.

Coding of Quantity and Accuracy

As explained earlier, the participants in all of the studies presented in this thesis saw the same video of a mock crime and responded to multiple questions about their memory for this stimulus. With the exception of Study 1, participants also made control decisions (i.e., decided whether to volunteer or withhold each answer). In order to assess quantity and accuracy using mixed effects modelling, responses were coded as 0 (incorrect) or 1 (correct). The criteria for what was classed as correct and incorrect depended on the type of outcome being assessed. When quantity is calculated as an aggregate statistic for traditional analyses, the number of correct answers that are volunteered is divided by the number of questions asked (Koriat & Goldsmith, 1994). In effect, withheld answers are treated as incorrect. Thus, when coding quantity for mixed effects modelling, correct answers that were volunteered were coded as 1 (correct) and all other responses (including correct responses that were withheld) were coded as 0 (incorrect). Accuracy can be calculated in two ways in traditional analyses. First, it can be calculated by dividing the number of correct answers (both volunteered and withheld) by the number of questions (Koriat & Goldsmith, 1994). This type of accuracy can be termed forced-report proportion correct but as I did not calculate proportions, it is referred to as response accuracy throughout the thesis. When coding response accuracy for mixed effects modelling, incorrect answers were coded as 0 (incorrect) and correct answers were

coded as 1 (correct), regardless of whether they were volunteered or withheld. The second way of calculating accuracy in traditional analyses is to divide the number of correct answers that are volunteered by the total number of volunteered answers, ignoring all questions for which answers are withheld (Koriat & Goldsmith, 1994). This type of accuracy can be termed free-report accuracy, though I refer to it simply as accuracy throughout the thesis. When coding for accuracy for mixed effects modelling, all withheld answers were excluded and coded correct answers that were volunteered were coded as 1 (correct), and incorrect answers that were volunteered were coded as 0 (incorrect).

Assessment of Monitoring and Response Bias

Experiments 2-5 explored the impact of particular manipulations on monitoring ability. Participants in these experiments had the opportunity to make control decisions (volunteer or withhold), and their answers were coded as either correct or incorrect. Traditionally, this type of data would be assessed using Type-2 Signal Detection Theory (SDT) measures whereby hit and false alarm rates are used to calculate a measure of discrimination (i.e., the extent to which correct details are volunteered and incorrect details are withheld) and a corresponding measure of response bias (i.e., tendency to volunteer information) (Higham, 2002). However, as explained earlier, calculation of these aggregate measures is problematic when performance is at the extremes (i.e., when hit or false alarm rates are 0% or 100%) and they do not take variance due to item into account. The mixed effects modelling I used is based on the work of Murayama et al. (2014), and while the approach assesses monitoring and response bias in a way that is analogous to Type-2 SDT measures, it avoids the limitations of the traditional by assessing all data points and accounting for random effects. The models presented in the thesis considered the extent to which control decisions (volunteer or withhold) were predicted by response accuracy (correct or incorrect) across different experimental manipulations.

Each model was constructed by initially entering control decisions (i.e., volunteer vs. withhold) as the outcome measure. After adding the random effects, experimental condition, response accuracy, and the interaction between experimental condition and response accuracy were entered as predictor variables. The experimental condition variable assesses response bias by considering whether the tendency to volunteer differs between the control condition and each experimental condition. The response accuracy variable assesses the extent to which the accuracy of a response predicts control decisions. Thus, it is a global measure of monitoring and gives no information about the extent to which monitoring differs between the control condition and each experimental condition. This information is obtained from the interaction. Throughout the thesis, the results of these analyses are presented in figures along with corresponding tables containing the coefficients for each experimental condition (i.e., response bias) and the interaction between experimental condition and response accuracy (i.e., monitoring). An example figure (Figure 1) and table (Table 2) from Experiment 2 are provided below.

As explained earlier, Experiment 2 considered whether instructions about mnemonic cues could improve monitoring and whether such instructions impact on the quantity and accuracy of eyewitness memory reports. In the figure, the predicted logs odds of volunteering is presented in the *y*-axis and experimental condition is presented on the *x*-axis. In this particular analysis, the clarity + reasoning, clarity + thoughts, and visual detail + effortfulness conditions were compared to the confidence condition. Monitoring is represented by the relative difference between the log odds of volunteering correct and incorrect responses, while the combined height of the correct and incorrect bars represents response bias. A larger difference between the correct and incorrect bars reflects superior monitoring and a higher combined height indicates that there is a stronger bias towards volunteering.

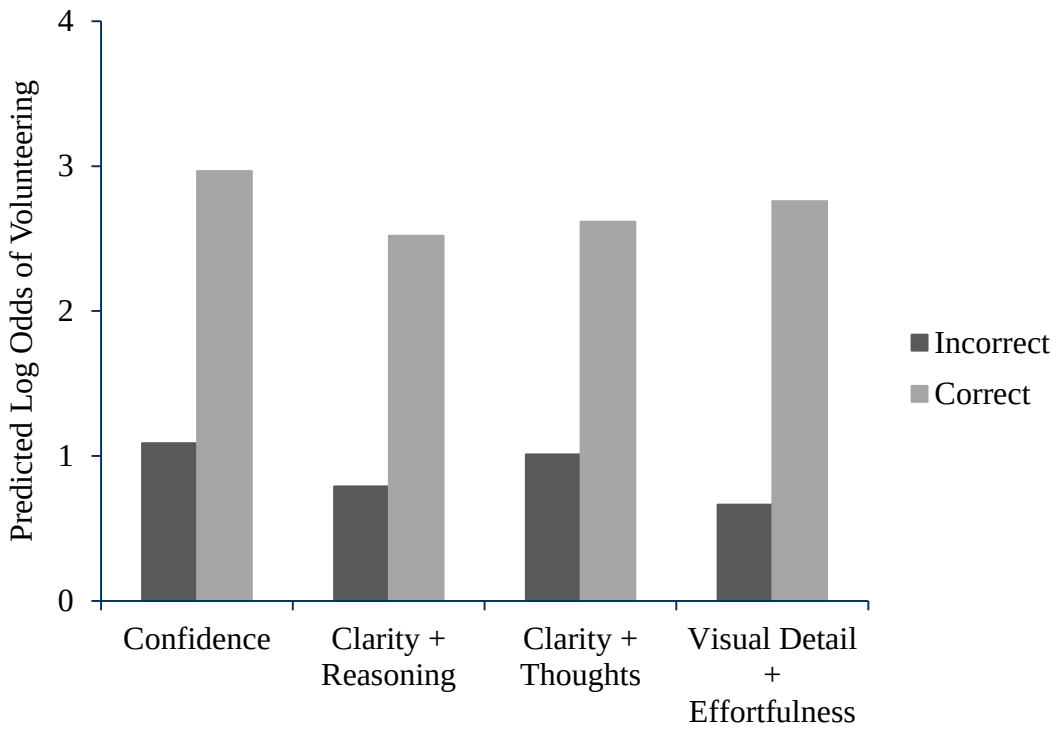

Figure 1. Predicted log odds of volunteering correct and incorrect answers as a function of the type of instructions provided in Experiment 2.

Figure 1 shows that the relative difference between the bars for correct and incorrect responses is similar across conditions. However, in comparison to the confidence condition, the difference is slightly smaller for the clarity + reasoning and clarity + thoughts conditions, and slightly larger for the visual detail + effortfulness condition. Consistent with this, the coefficients for monitoring presented in Table 2 were negative for the clarity + reasoning and clarity + thoughts conditions, and positive for the visual detail + effortfulness condition. Thus, Figure 1 and Table 2 show that the clarity + reasoning and clarity + thoughts conditions had less effective monitoring than the confidence condition, while the visual detail + effortfulness condition had more effective monitoring than the confidence condition. However, as the 95% CIs presented in Table 2 include zero, none of the differences were significant. Furthermore, a model containing the interaction between

experimental condition and response accuracy as a predictor variable did not fit the data significantly better than a model that did not contain the interaction between experimental condition and response accuracy as a predictor variable, χ^2 (3) = 1.51, p = .680.

Table 2

Fixed Effects Coefficients for Response Bias and Monitoring for the Clarity + Reasoning, Clarity + Thoughts, and Visual Detail + Effortfulness Conditions in Comparison to the Confidence Condition in Experiment 2

	Response bias		
Condition	*b*	SE_b	95% CI_b
Clarity + reasoning	-0.41	0.34	-1.07, 0.26
Clarity + thoughts	-0.39	0.34	-1.05, 0.28
Visual detail + effortfulness	-0.35	0.34	-1.02, 0.33
	Monitoring		
Clarity + reasoning	-0.15	0.40	-0.93, 0.64
Clarity + thoughts	0.22	0.41	-0.59, 1.02
Visual detail + effortfulness	-0.27	0.41	-1.08, 0.54

Figure 1 also shows that the combined height of the correct and incorrect bars is similar across conditions; though it is slightly lower for each of the experimental conditions than for the confidence condition. Consistent with this, all the coefficients for response bias presented in Table 2 were negative. Thus, Figure 1 and Table 2 show that participants were less likely to volunteer answers in each experimental condition than in the confidence condition. However, as the 95% CIs presented in Table 2 include zero,

none of the differences were significant. Furthermore, a model containing experimental condition as a predictor variable did not fit the data significantly better than a model that did not contain experimental condition as a predictor variable, χ^2 (3) = 1.81, p = .613.

Summary of Analytical Approaches

The data presented throughout this thesis were analysed using a logistic mixed effects modelling approach. This method of analysis was selected due to the risks associated with traditional approaches for categorical outcome measures, and the benefit of controlling for random variation between participants and items. Assessments of significance were made by considering whether model fit was significantly improved by adding the predictor variables, and by considering whether the confidence intervals of regression coefficients included zero.

CHAPTER 3 – MNEMONIC CUES AND RESPONSE ACCURACY

Study 1

In Chapter 1 I explained that it may be possible to improve monitoring by informing people of differences in the mnemonic cues associated with correct and incorrect memories. I argued that the source monitoring framework (Johnson et al., 1993) provides a description of the types of mnemonic cues which may be associated with correct and incorrect memories. Specifically, I proposed that mnemonic cues associated with memories created by an external stimulus may be associated with correct memories. Conversely, I proposed that mnemonic cues associated with memories created internally may be associated with incorrect memories. I also provided a summary of the literature which has explored the mnemonic cues outlined in the source monitoring framework. Although research within the literature has shown a consistent difference in the mnemonic cues associated with different kinds of memories, much of it considers memory errors that occur due to misinformation or deliberate imagination of events that did not occur, which are termed suggestion-dependant distortions (Mazzoni, 2002). These suggestion-dependant distortions have been studied at the expense of errors that occur due to the way memory operates which are termed naturally occurring distortions (Mazzoni, 2002). Furthermore, prior research has not considered whether witnesses already consider the mnemonic cues when monitoring the accuracy of their memories. Study 1 aimed to address the limitations of past research by (i) exploring whether the mnemonic cues outlined in the source monitoring framework could discriminate between naturally occurring correct and incorrect memories (i.e., memories not created via suggestion or deliberate imagination), and (ii) assessing whether the mnemonic cues were predictors of response accuracy after controlling for natural monitoring ability (i.e., the extent to which people can differentiate between correct and incorrect memories without any intervention).

The primary purpose of Study 1 was to identify mnemonic cues that could be used, in subsequent experiments, to improve monitoring.

I chose to focus on naturally occurring memory distortions because they have not been adequately in the literature, despite the fact that they are an important source of memory errors (Mazzoni, 2002). As noted earlier, previous research into mnemonic cues has often focussed on suggestion-dependent distortions. For example, Johnson et al. (1988) studied real and imagined autobiographical events, while Schooler et al. (1986) and Lane et al. (2007) studied memories for an event that was actually witnessed and misinformation about the witnessed event. Although the DRM task does constitute a type of natural memory distortion due to the unconscious nature of the memory errors that are produced (Mazzoni, 2002), the task is based on sematic memory rather than episodic memory. Furthermore, while some research has considered the relationship between mnemonic cues and naturally occurring correct and incorrect memories (Robinson, Johnson, & Herndon, 1997, Robinson et al., 2000), there are limitations to this work. Robinson and colleagues (1997, 2000) had participants recall details about a video and rate the extent to which they needed to reconstruct their memory for each detail, how vivid each detail was, how much effort was required to retrieve each detail, and how confident they were that each detail was correct. The results of these studies showed that memories rated as vivid were more likely to be correct, memories rated as involving reconstruction were more likely to be incorrect, and memories rated as being effortful to retrieve were more likely to be incorrect. These results are consistent with the findings of studies which have assessed suggestion-dependent memory distortions (as discussed in Chapter 1). However, Robinson and colleagues (1997, 2000) did not assess all of the mnemonic cues that have been outlined in the source monitoring framework. Thus, a main focus of Study

1 was to expand upon the work of Robinson and colleagues by assessing a larger variety of the mnemonic cues outlined in the source monitoring framework.

In keeping with the broad aim of identifying mnemonic cues that could be manipulated in subsequent experiments, I needed to determine the extent to which people spontaneously consider the mnemonic cues when monitoring response accuracy. As explained in Chapter 1, people usually engage in heuristic processes when determining the origin of a memory, though they are capable of using more systematic judgments to assist their decision making process (Johnson et al., 1993). I argued that it may be possible to improve monitoring by encouraging systematic processing. However, as some systematic processing of the mnemonic cues may be occurring naturally, it was important to assess the extent to which people engage in systematic assessment of the mnemonic cues without intervention. Natural monitoring ability is thought to be represented by confidence judgments which are known to predict response accuracy (Goldsmith & Koriat, 2008; Koriat & Goldsmith, 1996; Weber & Brewer, 2008). Therefore, confidence judgments should be based upon the heuristic processes that are used to judge memory accuracy, as well as any systematic processes that people usually engage in. However, much of the research into mnemonic cues has not measured confidence in addition to the mnemonic cues (Johnson et al., 1988; Lane et al., 2008, 2007; Norman & Schacter, 1997; Schooler et al., 1986). Furthermore, the only studies that have considered some of the mnemonic cues outlined in the source monitoring framework did not control for confidence when assessing the relationship between the mnemonic cues and response accuracy (Robinson, Johnson, & Robertson, 2000). Thus, previous research has not explored whether the mnemonic cues discriminate between different types of memories after controlling for natural monitoring ability. Consequently, it is possible that the mnemonic cues assessed in the literature form part of the natural monitoring processes of witnesses. If this is the case, any manipulation

of the mnemonic cues identified in the literature may not improve eyewitness monitoring. To address this gap in knowledge, confidence judgments were collected in Study 1 in addition to ratings of a variety of the mnemonic cues outlined in the source monitoring framework to determine whether any of the mnemonic cues predicted response accuracy to a greater extent than confidence alone.

Obtaining confidence judgments also allowed me to explore the basis of confidence judgments, though this was not the primary focus of the study. Currently, research into the basis of confidence judgments in eyewitness testimony is limited. However, there is some evidence which suggests that witnesses may consider ease of visualisation and subjective retrieval effort when judging how confident they are about the accuracy of their memories. As explained earlier, Robinson et al. (1997) had participants recall details about a video they had been shown and rate: (i) the extent to which they needed to reconstruct their memory for the item (as opposed to being able to visualise it easily), (ii) how much effort was required to retrieve each detail, and (iii) how confident they were that each detail was correct. The results showed that reconstruction and perceived retrieval effort were negatively related to confidence. Specifically, memories that were effortful to retrieve and memories that required reconstructive processing received lower confidence ratings. In a similar study, Robinson et al. (2000) had participants recall details about a video they had been shown and rate how vivid each detail was, how much effort was required to retrieve each detail, and how confident they were that each detail they provided was correct. The results showed that both memory vividness and perceived retrieval effort were significant predictors of confidence. Furthermore, the participant's estimates of retrieval fluency (i.e., estimate of the number of seconds taken to answer each question) also predicted confidence. Thus, confidence judgments may be based primarily on vividness, the extent

to which a memory is reconstructed/visualised, perceived retrieval effort, and perceived retrieval fluency.

In Study 1, participants watched a stimulus video depicting a mock bank robbery and answered a series of closed questions about it. In accordance with the source monitoring framework, I assessed three sensory characteristics (i.e., visual detail, clarity, and vagueness) and four cognitive processes (i.e., reasoning, thoughts, effortfulness, and perceived retrieval fluency) that were associated with the memories participants retrieved. I predicted that the extent to which memories were associated with visual detail and clarity would correlate positively with response accuracy. Conversely, I predicted that the extent to which memories were associated with vagueness, reasoning, thoughts, effortfulness, and perceived retrieval fluency would correlate negatively with response accuracy.

Method

Participants. Sixty-three participants[17] (53 females and 10 males) took part in the study for course credit or payment ($15). All had normal or corrected-to-normal vision and spoke English as their first language. Participants ranged in age from 17 to 48 (M = 20.49, SD = 4.93).

Materials.

Stimulus video. As my focus was on naturally occurring correct and incorrect memories, I needed a stimulus video that would elicit a sufficient number of incorrect details for analysis. I chose a video used by Tuckey and Brewer (2003) which has been

[17] Before data collection began I decided to recruit 60 participants and ended data collection in the week that I reached 60 participants. One participant did not follow instructions correctly, typing a 'don't know' response as their answer to 38.89% of the closed questions they answered. Another participant appeared to misunderstand 20% of the closed questions they answered. These responses were ignored in the analyses but the participants were not excluded.

found to produce memory errors due to the ambiguous stimuli incorporated within it (Tuckey & Brewer, 2003). As ambiguous stimuli are likely to be a common source of memory errors in real crimes (Tuckey & Brewer, 2003), the use of this video also strengthens the applied implications of my findings.

The video depicted a staged bank robbery with a duration of approximately 60 s. Two robbers, a male and a female, approached and entered a bank. The female robber's gender was made to appear ambiguous (i.e., no typical feminine characteristics such as long hair and body shape, were observable). The male went to the counter while the female stood off to the side pointing a bag at the counter as though it might contain a weapon. The male robber did not explicitly demand money but held his jacket pocket up to the cashier as though it might contain a gun. He also told the female cashier to hurry as she emptied the cash draw. The cashier handed the money to the male robber who tucked it away, though it was not clear whether it was put into a bag or his pocket due to the angle at which the scene was filmed. The robbers then left the bank and escaped. The male got onto a bus and the female escaped on foot.

Recall task. Participants were asked a series of closed questions about their memory for the video. For each robber, questions asked for a general physical description (i.e.., gender and build), a description of their clothing, a description of any weapons and/or bags, and a description of how each robber escaped. As the robbers wore different items of clothing (e.g., one wore a jacket and the other wore a jumper), I used yes/no filter questions such as 'Was the robber wearing a jumper?' and presented the same set of questions for both robbers. Including the filter questions also assisted in avoiding guess responses for items participants believed they did not see. When participants responded *yes* to a filter question, they were presented with either one or two closed questions that requested a description of the item in question. Thus, although there were 14 filter

questions and 34 closed questions in total (see Appendix A),[18] the number of closed questions participants were presented with depended partially on responses to the filter questions. For example, one participant may have indicated that the robber by the counter was wearing a disguise and subsequently been asked about the colour and type of disguise worn. Conversely, another participant may have indicated that the robber by the counter was not wearing a disguise and, therefore, would not have been presented with the two follow up questions. Assuming all other filter questions were answered in the same way by both participants, the first participant would have answered two additional questions.

As the purpose of this study was to identify mnemonic cues that could be used to improve eyewitness monitoring, it was important to collect data for all memories that could be retrieved, including those that participants were not very confident about. Thus, participants completed the closed questions in a forced-report format (i.e., they did not have the option of withholding answers to the closed questions). Although this may have resulted in some guesses, the filter questions are likely to have prevented participants from creating answers about things they believed they did not see.

A scoring guide was constructed to code the participants' answers as either correct or incorrect. A mean number of 22.48 (*SD* = 2.65) responses were coded as either correct or incorrect per participant and coding was completed by myself and an independent rater using all of the data. There was an acceptable level of interrater agreement, $\kappa = .83$ [.80, .85]. Although most of the incorrect answers appeared to be genuine erroneous memories, some seemed to result from confusion. For example, some participants appeared to confuse the robbers entirely, or provided information in the wrong question (see Appendix

[18] Only data from the closed questions is included in the analyses presented throughout the thesis.

B for an explanation of how these responses were coded).[19] Two data sets were constructed; one that included all of the data and another that ignored any items that involved some form of coder interpretation (e.g., items where participants were judged to have confused the robbers or to have provided information in the wrong question). All analyses were run using both data sets. The results presented are for the full data set. However, where differences emerged, the results for data set which ignored responses requiring coder interpretations are presented as footnotes.

Confidence scale. I used a confidence scale ranging from 0% to 100% with 10% intervals (i.e., 0%, 10%, 20%, etc.) to measure the participants' confidence in their answers.

Mnemonic cue rating scales. I used rating scales to measure the mnemonic cues because they are more precise than open-ended questions, and allow smaller differences to be detected between correct and incorrect information (Lampinen, Neuschatz, & Payne, 1997). However, rating scales may also prompt people to search for information which they would not normally consider (Lampinen et al., 1997). Although Lampinen et al. (1997) view this as a limitation, it is actually an advantage in this case because my aim was to identify information that people have access to, but do not consider spontaneously when monitoring the accuracy of their memories. Thus, it is essential that the rating scales prompt people to think about information that they may not spontaneously consider, rather than simply indexing the information people spontaneously draw upon.

The rating scales were based on the source monitoring framework, and were designed to elicit information about the sensory characteristics and cognitive processes

[19] Sometimes, participants also appeared to misunderstand the question. For example, it is clear that the participant who responded 'brown' to the question 'What was the general build of the robber by the counter?', did not understand the question. Answers where participants appeared to misunderstand the question were ignored because they do not constitute memory errors.

associated with the participants' memories. I used seven ratings scales, each starting with the stem "My memory for the item", and completed with: (i) is effortful, (ii) involves reasoning, (iii) includes many thoughts, (iv) is clear, (v) is visually detailed, (vi) is vague, and (vii) came to mind quickly.[20] As access to information about cognitive processes decays quickly (Ericsson & Simon, 1980), the scales measuring perception of cognitive processes were presented first, followed by the scales measuring sensory characteristics and, finally the perceived retrieval fluency scale. The ratings were made on a five-point scale ranging from 0 (not at all) to 4 (very much). This five-point scale was selected because scales with fewer than five points are unreliable and people have difficulty using scales with more than seven points (Thomas, 2004).

Procedure. After giving informed consent, participants were shown the video, followed by a 10-minute delay in which they completed mazes of increasing difficulty. Participants then answered the questions on a computer. Filter questions were answered by clicking either a *yes* or a *no* button. Regardless of which button was selected, participants rated how confident they were in their response. After rating confidence, they proceeded to the next question if they had selected *no* in response to the filter question, or proceeded to a closed follow-up question if they had selected *yes* in response to the filter question. For each closed question, a text box was provided in which participants typed their answer. Participants then completed the seven mnemonic cues ratings and a confidence rating. The mnemonic cue ratings were completed after each question rather than after all questions because the characteristics of memories are strongest at the time of retrieval, particularly any cognitive processes involved (Ericsson & Simon, 1980). Filter questions were

[20] No additional information was provided about these anchors. Additional explanation of such anchors does not appear to be common in the literature given that no mention is made of such explanations in the methodology (see Johnson et al., 1988 for an example).

proceeded by either one or two closed questions. When there were two closed questions, participants answered the first question, rated the mnemonic cues and confidence for their answer to the first question, then repeated this process for the second closed question. Participants were instructed to answer all questions.

For each question in the recall task I measured only: (i) the participant's response, (ii) their confidence in the response, (iii) the rating they gave for each mnemonic cue (closed questions only), and (iv) the reaction time for each of these responses (which were not analysed). There were no manipulations.

Results

The relationship between mnemonic cues and confidence. Initially, I was interested in conducting discriminant validation to determine whether any of the mnemonic cues were measuring the same constructs. Such mnemonic cues would not have needed to be analysed separately. Furthermore, as the purpose of Study 1 was to identify mnemonic cues that could be used to improve eyewitness monitoring (which is represented by confidence), I also examined the extent to which each mnemonic cue was related to confidence. Mnemonic cues that are strongly related to confidence are likely to be measuring confidence or a construct upon which confidence is based, though they may not be the only basis of confidence judgments. Mnemonic cues that are strongly related to confidence would be unlikely to improve monitoring if they were manipulated because their relationship with confidence would indicate that they are considered spontaneously when witnesses judge the accuracy of their memories.

I conducted a series of correlations to test the relationships between the mnemonic cues and confidence (Table 3). In accordance with recommendations in the literature (Gelman & Hill, 2006), a cut-off of .80 was used to determine whether the variables were measuring the same constructs because this is the level at which multi-collinearity

Table 3

Correlations Between the Seven Mnemonic Cues and Confidence

		Effortfulness	Reasoning	Thoughts	Clarity	Visual detail	Vagueness	Perceived retrieval fluency	Confidence
Effortfulness	*r*		.63	.62	-.18	-.14	.30	-.21	-.20
	95% CI		[.60, .66]	[.59, .65]	[-.23, -.13]	[-.19, -.09]	[.25, .35]	[-.26, -.16]	[-.25, -.15]
Reasoning	*r*			.66	-.08	-.06	.24	-.08	-.11
	95% CI			[.63, .69]	[-.13, -.03]	[-.11, -.00]	[.19, .28]	[-.13, -.03]	[-.16, -.06]
Thoughts	*r*				-.06	-.01	.21	-.03	-.08
	95% CI				[-.11, -.01]	[-.06, .04]	[.16, .26]	[-.08, .02]	[-.13, -.03]
Clarity	*r*					.85	-.68	.76	.76
	95% CI					[.84, .87]	[-.70, -.65]	[.74, .78]	[.74, .78]
Visual detail	*r*						-.67	.72	.75
	95% CI						[-.70, -.64]	[.70, .74]	[.72, .77]
Vagueness	*r*							-.58	-.67
	95% CI							[-.62, -.55]	[-.69, -.64]
Perceived retrieval fluency	*r*								.71
	95% CI								[.68, .73]

Note. *df* = 1414 and *r* is significant at the .05 level when CI does not include 0.

becomes a problem. Variables with a correlation of below .80 are unlikely to be completely redundant with one another. With the exception of the correlation between clarity and visual detail, all correlations were below .80 which suggested that most of the mnemonic cues were measuring different constructs from each other and from confidence. However, as the correlations were unable to account for the mixed-effects structure of the data (i.e., the repetition of the ratings across questions and participants), it is possible that the relationships between the variables were overestimated. For example, it is possible that questions which elicited higher clarity ratings on average also elicited high visual detail ratings on average, making it appear as though clarity and visual detail were more closely related than they really are. Due to possibility of the inflated correlations, and because the correlation between clarity and visual detail was only just above the .80 cut-off (i.e., .85), I decided to treat clarity and visual detail as separate variables in all subsequent analyses.

The relationship between mnemonic cues and response accuracy. As the primary aim of Study 1 was to identify mnemonic cues that could be used to improve eyewitness monitoring in subsequent experiments, I examined whether the mnemonic cues predicted response accuracy to a greater extent than can be achieved naturally. According to Koriat and Goldsmith's (1996) metamemory framework, confidence judgments (i.e., the assessments of the likely accuracy of memory) are thought to represent people's natural monitoring ability. Thus, I controlled for the effect of confidence in the analyses. The data suggested that this was a sensible strategy given that confidence was a significant predictor of response accuracy, $b = 0.29$, $SE_b = 0.08$, [0.14, 0.44]. This means that for every 10% increase in confidence, the log odds of an accurate response increased by .29.

Seven logistic mixed effects models were constructed with response accuracy entered as the outcome measure. Confidence was always added as the first predictor, followed by one

of the mnemonic cues. Thus, the analyses assessed the extent to which the mnemonic cues predicted response accuracy in addition to what is already predicted by the confidence judgment. Each model was compared to a model that included only confidence and random intercepts for participant and question (see Table 4 for model fit statistics and Table 5 for coefficients). The results showed that model fit was significantly improved

Table 4

Model Fit Statistics for Logistic Mixed Effects Models Predicting Accuracy from the Mnemonic Cues after Controlling for Confidence in Study l

Fixed effect	$\chi^2(1)$	*p*
Effortfulness	1.10	.158
Reasoning	5.66	.017
Thoughts	5.55	.018
Clarity	5.69	.017
Visual detail	0.71	.399
Vagueness	4.66	.031
Perceived retrieval fluency	3.71	.054

Table 5

Fixed Effect Coefficients for Logistic Mixed Effects Models Predicting Accuracy from the Mnemonic Cues after Controlling for Confidence in Study 1

Fixed effect	*b*	SE_b	95% CI[b]
Effortfulness	-0.20	0.07	-0.34, -0.07
Reasoning	-0.24	0.07	-0.39, -0.10
Thoughts	-0.23	0.08	-0.38, -0.08
Clarity	0.53	0.08	0.38, 0.68
Visual detail	0.45	0.08	0.29, 0.60
Vagueness	-0.44	0.07	-0.59, -0.30
Perceived retrieval fluency	0.52	0.09	0.35, 0.69

with the addition of reasoning, thoughts, clarity[21] and vagueness, and that each of these variables was a significant predictor of response accuracy. The coefficients revealed that reasoning, thoughts, and vagueness were negative predictors of response accuracy while clarity was a positive predictor of response accuracy. Although model fit was not

[21] For the data set that ignored responses involving coder interpretations, clarity was not a significant predictor of response accuracy after controlling for confidence, $b = 0.20$, $SE_b = 0.11$, [-0.02, 0.42], and adding it to the model did not significantly increase fit, $\chi^2(1) = 3.06$, $p = .080$. This reduction in the coefficient is likely due to the fact that responses were often removed when participants confused the robbers. Such memories are likely to be poor and also lack clarity, meaning that their removal erases a subset of items for which clarity is very diagnostic of response accuracy. Therefore, excluding these items from analysis would have truncated variability (as some of the incorrect items most strongly related to clarity were removed), limited potential covariance, and resulted in the true effect being masked. For this reason, the analyses which used the full data set are likely to be more reliable and as a result my conclusions are based on those results.

significantly improved with the addition of perceived retrieval fluency, the coefficient showed that it was a significant predictor of response accuracy.[22] Specifically, it was a positive predictor of response accuracy. The fit of the model was not significantly improved with the addition of effortfulness or visual detail, and these mnemonic cues were not significant predictors of response accuracy after controlling for confidence.

I explored the findings for effortfulness and visual detail further to determine whether these mnemonic cues were completely unrelated to response accuracy. Specifically, I examined whether effortfulness and visual detail were significant predictors of response accuracy without controlling for confidence. Two new logistic mixed effects models were constructed (one for effortfulness and another for visual detail) that did not include confidence as a predictor variable. Model fit was significantly improved with addition of effortfulness, χ^2 (1) = 8.45, p = 0.33, and visual detail, χ^2 (1) = 31.05, p < .001. Furthermore, the coefficients showed that effortfulness, b = -0.20, SE_b = 0.07, [-0.34, -0.07], and visual detail, b = 0.45, SE_b = 0.08, [0.29, 0.60], were significant predictors of response accuracy with effortfulness being a negative predictor and visual detail a positive predictor. Thus, the findings showed that effortfulness and visual detail are related to

[22] For the data set that ignored responses involving coder interpretations, perceived retrieval fluency was not a significant predictor of response accuracy after controlling for confidence, b = 0.15, SE_b = 0.12, [-0.08, 0.37], and adding it to the model did not significantly increase fit, χ^2 (1) = 0.12, p = .217. As explained in the previous footnote, this reduction in the coefficient is likely due to the fact that responses were often removed when participants confused the robbers. Such memories are likely to be poor and also take longer to retrieve, meaning that their removal erases a subset of items for which perceived retrieval fluency is very diagnostic of response accuracy. Therefore, excluding these items from analysis would have truncated variability (as some of the incorrect items most strongly related to perceived retrieval fluency were removed), limited potential covariance, and resulted in the true effect being masked. For this reason, the analyses which used the full data set are likely to be more reliable and as a result my conclusions are based on those results.

response accuracy, though they do no uniquely predict response accuracy when confidence is taken into account.

Combinations of the mnemonic cues and their prediction of response accuracy. In addition to examining each mnemonic cue separately, I also explored whether there was a set of mnemonic cues that was most efficient at predicting response accuracy. Specifically, I examined whether any combination of the mnemonic cues significantly improved the fit of the model. In these analyses, the five mnemonic cues identified as significant predictors of accuracy were used and they were ordered from strongest to weakest according to absolute *b* value: clarity, perceived retrieval fluency, vagueness, reasoning, and thoughts. I began the analyses by creating a logistic mixed effects model containing the mnemonic cues with the first and second strongest *b* values (i.e., clarity and perceived retrieval fluency). This model was compared to the two simpler models that contained only one mnemonic cue, confidence, and the random effects. This process was repeated until all possible pairs had been considered. A pair was considered to be better when model fit was improved compared to both of the simpler models, and both of the mnemonic cues were significant predictors of response accuracy (see Table 6 for model fit statistics). Only two pairs significantly improved model fit over both simpler models: clarity + reasoning and clarity + thoughts. Furthermore, in each of these models both mnemonic cues were significant predictors of response accuracy (see Table 7 for coefficients). A model containing all three of these mnemonic cues was not significantly better than the clarity + reasoning model, χ^2 (1) = 1.35, p = .246, or the clarity + thoughts model, χ^2 (1) = 1.32, p = .251. Thus, the models that combined the clarity cue with either the reasoning cue or the thoughts cue provided the best fit to the data, optimising the prediction of response accuracy.

Table 6

Model Fit Statistics for Logistic Mixed Effects Models Predicting Accuracy from Mnemonic Cue Pairs in Study 1

	Mnemonic cue pairs							
	Clarity & perceived retrieval fluency		Clarity & vagueness		Clarity & reasoning		Clarity & thoughts	
Single mnemonic cue	$\chi^2(1)$	*p*	$\chi^2(1)$	*p*	$\chi^2(1)$	*p*	$\chi^2(1)$	*p*
Clarity	0.98	.322	2.02	.155	5.30	.021	5.33	.021
Perceived retrieval fluency	2.97	.085						
Vagueness			3.06	.080				
Reasoning					5.34	.021		
Thoughts							5.47	.019

Table 7

Fixed Effect Coefficients for Logistic Mixed Effects Models Predicting Accuracy from Mnemonic Cue Pairs in Study 1

	Clarity + reasoning		
Fixed Effect	*b*	SE_b	95% CI^b
Clarity	0.25	0.10	0.05, 0.45
Reasoning	-0.17	0.07	-0.32, -0.03
	Clarity + thoughts		
Clarity	0.25	0.10	0.05, 0.45
Thoughts	-0.18	0.08	-0.33, -0.03

Discussion

My primary aim in Study 1 was to identify mnemonic cues that could be used to improve monitoring in subsequent experiments. The results revealed that five mnemonic cues may have been able to help people monitor the accuracy of their memory for an episodic event more efficiently: reasoning, thoughts, clarity, vagueness, and perceived retrieval fluency. Furthermore, the findings showed that it may have been possible to affect the greatest change in monitoring by targeting the clarity cue and either the reasoning cue or the thoughts cue in subsequent experiments. However, it appeared that targeting the effortfulness and visual detail cues in subsequent experiments may not have a meaningful impact on monitoring because it appears that people already utilise these mnemonic cues when judging the accuracy of their memories. Thus, encouraging people to consider these mnemonic cues is unlikely to impact on their monitoring ability.

The results were consistent with the idea that the mnemonic cues outlined in the source monitoring framework can discriminate between naturally occurring correct and incorrect memories. In Chapter 1 I proposed that mnemonic cues associated with externally derived memories may be associated with correct memories while mnemonic cues associated with internally derived memories may be associated with incorrect memories. In support of this argument, the results showed that reasoning, thoughts, vagueness, perceived retrieval fluency, and effortfulness were negative predictors of response accuracy, while clarity and visual detail were positive predictors of response accuracy. These results are consistent with past research which has focussed on memory distortions produced by experimental manipulations such as misinformation or task content (Johnson et al., 1988; Lane et al., 2008, 2007; Norman & Schacter, 1997; Schooler et al., 1986), and with the work of Robinson and colleagues (1997, 2000).

Study 1 also extended upon previous research as the results showed that many of the mnemonic cues outlined in the source monitoring framework predict response accuracy after natural monitoring ability (i.e., confidence) is taken into account. As explained in the introduction of this chapter, previous research that has assessed the relationship between mnemonic cues and response accuracy has either not measured confidence (Johnson et al., 1988; Lane et al., 2008, 2007; Norman & Schacter, 1997; Schooler et al., 1986) or has not controlled for it when analysing the data (Robinson et al., 1997, 2000). Study 1 addressed both of these limitations and the results revealed that five of the mnemonic cues were significant predictors of response accuracy after controlling for confidence: reasoning, thoughts, clarity, vagueness, and perceived retrieval fluency. This indicated that people do not fully utilise these mnemonic cues when they monitor the accuracy of their memories. Therefore, it appeared that manipulating these mnemonic cues in subsequent studies could lead to improved monitoring and consequently allow people to maximise the quantity and

accuracy of their eyewitness memory reports. However, the findings also indicated that it may not be necessary to target all of these mnemonic cues in order to improve monitoring. Specifically, the findings indicated that encouraging people to make better use of the clarity cue in combination with either the reasoning cue or the thoughts cue were potentially the most efficient ways of improving monitoring.

The results of Study 1 also provided information regarding the basis of confidence judgments in eyewitness testimony. The fact that visual detail and effortfulness were no longer significant predictors of response accuracy after confidence was taken into account suggested that these mnemonic cues may inform eyewitness confidence judgments. This is consistent with the correlational evidence supplied by Robinson and colleagues (1997, 2000) which showed that confidence was negatively related to retrieval effort and positively related to vividness. However, it must be noted that while I observed a strong correlation between visual detail and confidence, the correlation between confidence and effortfulness was weak. If effortfulness is one of the mnemonic cues that witnesses use to judge confidence, I would have expected it to have a stronger correlation with confidence. Indeed, Robinson et al. (1997) observed a correlation of -.73, much higher than the -.20 correlation I observed. However, their correlation was calculated within-subjects rather than collapsed across participants and items as I have done. When calculated within-

subjects, the correlation between confidence and effortfulness was -.76,[23] consistent with Robinson et al.'s (1997) finding. Thus, my findings provided further evidence that the effortfulness and visual detail of memories informs confidence judgments.

The fact that reasoning, thoughts, clarity, vagueness, and perceived retrieval fluency remained significant predictors of response accuracy after controlling for confidence suggested that these mnemonic cue ratings are not completely captured within confidence judgments. The finding for clarity was somewhat surprising because it seems at odds with the result I observed for visual detail and the prior research conducted by Robinson and colleagues (1997, 2000) regarding vividness. I would have expected the clarity of a memory to be highly associated with its visual detail and vividness, meaning that clarity should not have remained a significant predictor of response accuracy after confidence was taken into account. The extent to which witnesses use clarity during monitoring is explored further in Experiment 2. The fact that perceived retrieval fluency remained a significant predictor of response accuracy after controlling for confidence also seems to contradict Robinson et al.'s (2000) finding that confidence judgments are based on estimates of retrieval fluency. However, this discrepancy may be explained by a key methodological difference between Study 1 and Robinson et al.'s (2000) study.

[23] This disparity is most likely due to the fact that different participants used the effortfulness scale in quite different ways. For example, one participant may use a 2 to indicate a relatively high level effortfulness and 0 to indicate a relatively low level of effortfulness, whereas another participant may use 4 to indicate a relatively high level of effortfulness and 2 to indicate a relatively low level of effortfulness. By collapsing these ratings across participants, this variation would have artificially reduced the relationship between effortfulness and confidence. It is important to note that when a random intercept is included, mixed effects models can account for the difference in scale usage while still considering all of the data. Thus, mixed effects models provide the best possible analysis of data like this.

Specifically, participants in my study were asked to rate how quickly each memory came to mind, while Robinson et al.'s (2000) participants provided an estimate of the number of seconds taken to answer each question. It could be that specific estimates provide a better basis for confidence judgments.

In conclusion, Study 1 provided evidence that naturally occurring correct and incorrect memories can be distinguished by their association with a variety of mnemonic cues. It also appeared as though people do not fully utilise some of these mnemonic cues (i.e., reasoning, thoughts, clarity, vagueness, and perceived retrieval fluency) when they monitor the accuracy of their memories. Therefore, based on these results, it seemed that manipulating these mnemonic cues in subsequent experiments could result in improved monitoring. However, the findings also indicated that the greatest benefits in monitoring might be produced by targeting just two of the mnemonic cues: clarity and reasoning or clarity and thoughts.

CHAPTER 4 – IMPROVING MONITORING: MANIPULATING KNOWLEDGE OF MNEONIC CUES

In two experiments, I manipulated knowledge of the mnemonic cues in an attempt to improve monitoring and maximise the quantity and accuracy of memory reports about an episodic event. In contrast to Study 1, Experiments 2 and 3 used free-report questioning procedures which gave participants the opportunity to regulate the information contained in their eyewitness memory reports.[24] Specifically, for each question, participants made a control decision whereby they decided whether to volunteer or withhold their response. Knowledge of the mnemonic cues was manipulated by providing participants with information about particular mnemonic cues and encouraging them to consider these mnemonic cues when making control decisions. Similar manipulations have been found to improve performance on source memory (Lane et al., 2007) and recognition tests (Lane et al., 2008), as explained in Chapter 1. The aim of these experiments was to maximise type-2 discriminability; indicated by an increase in the amount of correct details that were volunteered (i.e., hits) and/or a decrease in the amount of incorrect details that were volunteered (i.e., false alarms). It was expected that maximising discriminability would allow people to maximise the accuracy of their eyewitness memory reports with minimal cost to quantity.

[24] The fact that participants were not given the explicit option of withholding responses does not mean that they were completely unable to regulate their memory reports in some other way (e.g., by giving memories they would have withheld a very low confidence rating).

Experiment 2

The design of Experiment 2[25] was based on the results of Study 1 which showed that the clarity, reasoning, thoughts, vagueness, and perceived retrieval fluency cues were significant predictors of response accuracy after controlling for confidence. Study 1 also revealed that combining the clarity cue with either the reasoning cue or the thoughts cue was found to be the most efficient way of predicting response accuracy. This suggested that providing witnesses with instructions about one of these mnemonic cue pairs may be the best way to improve monitoring. As Study 1 provided no basis for choosing between these pairs, one experimental condition received instructions about the clarity and reasoning cues, and the other received instructions about the clarity and thoughts cues.

Two control conditions were also included in Experiment 2. The first control condition received information about confidence because such instructions were not expected to alter performance (i.e., type-2 discrimination, quantity, or accuracy). Recall that Koriat and Goldsmith's (1996) metamemory framework proposes that people use confidence to judge response accuracy, and that Study 1 provided support for this proposition given that confidence was found to be a significant predictor of response accuracy. Therefore, I did not expect monitoring judgments or control decisions to be affected by telling people to consider something they already use without being encouraged. The second control condition received instructions about the visual detail and effortfulness cues because these cues were not found to be significant predictors of response accuracy after controlling for confidence in Study 1. Thus, I did not expect instructions about these mnemonic cues to alter performance. Including this control condition allowed me to examine whether any differences in performance between the

[25] Although this was the first experiment, it was the second study I conducted. Thus to ensure that no two studies shared the same number I chose to call my second study Experiment 2.

confidence and experimental conditions were the result of (i) the consideration of potentially useful mnemonic cues (i.e., clarity and reasoning or thoughts) when judging response accuracy and making control decisions, or (ii) receiving information about any mnemonic cues that prompt more careful consideration of response accuracy and control decisions.

I predicted that the experimental conditions would have better monitoring (type-2 discriminability) than each of the control conditions. However, the impact of any improvements in monitoring on quantity and accuracy could manifest in different ways. For accuracy to be improved the proportion of volunteered information that is correct must be increased. This can be achieved by (i) increasing the amount of correct information being volunteered, (ii) reducing the amount of incorrect information being volunteered, or (iii) a combination of both. Each of these options constitutes an improvement in monitoring. Thus, if the experimental conditions exhibited better monitoring than each of the control conditions, accuracy was also expected to be higher for the experimental conditions than for each of the control conditions. However, to avoid reducing quantity, the proportion of retrieved information that is both correct and volunteered must be maximised. This means that I only expected quantity to be higher for the experimental conditions if they volunteered a higher amount of correct information than the control conditions.

Method

Participants. One-hundred and twenty participants[26] (85 females and 35 males) took part in the study for course credit or payment ($15). All had normal or corrected-to-

[26] Before data collection began I decided to recruit 120 participants, 30 per condition. Initially, I recruited 35 participants to test the manipulation. After determining that the manipulation was working satisfactorily

normal vision and spoke English as their first language. Participants ranged in age from 18 to 61 (M = 22.66, SD = 6.07).

Materials.

Stimulus video and recall task. Participants viewed the same video and completed the same questions that were used in Study 1. However, several small changes were made to the questions due to the responding irregularities observed in Study 1. To avoid confusion over the term 'disguise', the disguise questions were presented after the clothing questions. An extra instruction was also included before participants answered the questions to reduce the likelihood of participants providing information in the wrong question. Specifically, participants were informed that there would be several questions about the top/s possibly worn by each robber. They were told that there would be questions asking whether each robber was wearing a jacket, shirt and/or jumper and were asked to respond in the appropriate question. They were given the following example,

'…if you remember that one of the robbers was wearing a jumper, but not a jacket, respond *No* to the question that asks if they were wearing a jacket and *yes* to the question that asks if they were wearing a jumper.'

Monitoring instructions. Participants in the clarity + reasoning, clarity + thoughts, and visual detail + effortfulness conditions were told that research had identified two memory characteristics which help distinguish between correct and incorrect memories. They were encouraged to consider this information when judging the correctness of their memory and deciding what to volunteer. Participants in the clarity + reasoning and clarity + thoughts conditions both received information about the clarity cue. They were told that

---

(based on analysis of the manipulation check questions) I collected the remaining 85 participants and ended data collection in the week that I reached 120 participants.

clear memories are likely to be correct and should be volunteered. Participants in the clarity + reasoning condition were also told that memories involving reasoning are likely to be incorrect and should be withheld, while participants in the clarity + thoughts condition were told that memories including many thoughts are likely to be incorrect and should be withheld. Participants in the visual detail + effortfulness condition received information about the visual detail and effortfulness cues. Specifically, they were told that visually detailed memories are likely to be correct and should be volunteered, and that effortful memories are likely to be incorrect and should be withheld. Participants in the confidence condition were asked to rely on confidence when judging the correctness of their memory and deciding what to volunteer. They were told that memories witnesses are highly confident about are likely to be correct and should be volunteered.

Confidence scale. As in Study 1, confidence was measured using a 0% to 100% scale with 10% intervals (i.e., 0%, 10%, 20%, etc.).

Mnemonic cue rating scales. For each closed question, participants completed ratings for the two mnemonic cues they were instructed to consider when judging the correctness of their memory and making control decisions. The ratings scales were the same as those used in Study 1. Thus, participants were asked to rate the extent to which each of their answers was associated with the mnemonic cues on a five-point scale ranging from 0 (not at all) to 4 (very much).

Although research has not found any advantage of obtaining ratings in addition to providing mnemonic cue instructions (Lane et al., 2008, 2007), I believed it was important to obtain these ratings for two reasons. First, the ratings constitute a second part of the manipulation as they served as a reminder to the participants that they should consider the mnemonic cues throughout the recall task. Second, the ratings allowed me to assess the replicability of my Study 1 results. Third, in the event of ineffective manipulations, the

ratings allowed me to test whether (i) the instructions were ineffective because participants did not use the mnemonic cues they were instructed to consider when making their control decisions, or (ii) the mnemonic cues were simply unreliable indexes of response accuracy in these circumstances.

Manipulation check questions. I included 14 questions after the closed questions as a manipulation check to assess whether participants remembered and understood the instructions. The first two questions[27] were in a multiple choice format and asked the participants (i) what information they should use when judging whether their memory is correct and (ii) what information they should use when deciding whether to volunteer or withhold information. They were instructed to select two responses from the following options: (i) level of confidence in the memory, (ii) whether the memory is clear, (iii) whether the memory involves reasoning, (iv) whether the memory includes thoughts, (v) whether the memory is visually detailed, and (vi) whether the memory is effortful. The next six questions asked participants to rate how likely memories were to be correct when they were associated with high confidence or an experience of each mnemonic cue on a scale ranging from 0 (likely to be incorrect) to 4 (likely to be correct) scale. The final six questions asked whether participants should volunteer or withhold an answer if: (i) they were highly confident in their memory, (ii) their memory was clear, (iii) their memory involved reasoning, (iv) their memory included a lot of thoughts, (v) their memory was visually detailed, and (vi) their memory was effortful. All participants were presented with the full set of manipulation check questions regardless of condition.

Procedure. After giving informed consent, participants were randomly assigned to one of four levels of the only manipulation: clarity + reasoning, clarity + thoughts, visual

[27] These questions were added after pilot testing. Thus, data for these questions are only available for 85 of the 120 participants.

detail + effortfulness or confidence (n = 30 in each condition). Participants were then shown the stimulus video and completed the 10-min maze filler task that was used in Study 1. All participants completed the experiment on a computer.

After the filler task, participants received their monitoring instructions (the exact wording of all instructions is provided in Appendix C). These instructions were repeated after a practice filter and closed question (also included in Appendix C), and participants then completed the recall task. When presented with filter questions, participants had two response options, *yes* and *no*. When they said *yes*, they were presented with a closed follow-up question and typed an answer in the text box provided. Participants in the clarity + reasoning, clarity + thoughts, and visual detail + effortfulness conditions then rated the two mnemonic cues that were relevant to their condition. All participants then rated how confident they were that their answer was correct. Finally, they were asked whether they wanted to include the answer in their eyewitness memory report by selecting one of two response options: *Submit Answer* or *Withhold Answer*. When all the recall questions were complete, participants responded to the manipulation check questions.

For each question in the recall task I measured only: (i) the participant's response, (ii) their confidence in the answer, (iii) their mnemonic cue and ratings (closed questions only), (iv) whether they chose to volunteer or withhold the answer (closed questions only), and (v) the reaction time for each of these responses. Their responses to the manipulation check questions, and the reaction times of these responses were also recorded (none of the reaction time data were analysed).

Results

The relationship between mnemonic cues and response accuracy. To assess the replicability of my findings from Study 1 I examined whether the mnemonic cues

predicted response accuracy. A separate logistic mixed effects model was constructed for each mnemonic cue and response accuracy was used as the outcome measure because it allows all of the participants' answers to be examined. I did not control for confidence in these analyses because the instructions appeared to be having an impact on confidence judgments. This was illustrated by the fact that the correlations between confidence and each of the mnemonic cues that were measured in Experiment 2 (i.e., clarity, thoughts, visual detail, and effortfulness) were stronger than those observed in Study 1 (Table 8). Furthermore, the CI's for these mnemonic cues did not overlap between Study 1 and Experiment 2, suggesting that the differences were meaningful. However, the correlation between confidence and reasoning was very similar across both studies and there was a high degree of overlap in the CI's which suggested a negligible difference. Thus, with the exception of reasoning, participants seemed to have used the mnemonic cues when making confidence judgments. Therefore, controlling for confidence when assessing whether the mnemonic cues predicted response accuracy may have masked some of the effects.

The results for the logistic mixed effects models revealed that model fit was significantly improved with the addition of clarity, visual detail, effortfulness, and thoughts, but not with the addition of reasoning (Table 9). Furthermore, clarity, visual detail, effortfulness, and thoughts were significant predictors of response accuracy, while reasoning was not (Table 10). These findings replicated my Study 1 results for the clarity, visual detail, effortfulness, and thoughts cues, but not the results for the reasoning cue. This suggested that reasoning may not be a reliable predictor of response accuracy.

Table 8

Comparison of the Correlations between Confidence and Five of the Mnemonic Cues for Study 1 and Experiment 2

		Study 1	Experiment 2
Clarity	*r*	.76	.90
	95% CI	[.74, .79]	[.89, .91]
Reasoning	*r*	-.12	-.13
	95% CI	[-.17, -.07]	[-.20, -.06]
Thoughts	*r*	-.09	-.33
	95% CI	[-.14, -.04]	[-.39, -.23]
Visual detail	*r*	.74	.88
	95% CI	[.72, .77]	[.86, .89]
Effortfulness	*r*	-.21	-.48
	95% CI	[-.26, -.16]	[-.54, -.42]

Note. All *r* values were significant with $p < .001$. The correlations between confidence and each mnemonic cue in Experiment 2 are for a subset of the total sample as not all participants completed all mnemonic cue ratings. Participants in the clarity + reasoning and clarity + thoughts conditions contributed to the correlation for clarity, participants in the clarity + reasoning condition contributed to the correlation for reasoning, participants in the clarity + thoughts condition contributed to the correlation for thoughts, and participants in the visual detail + effortfulness conditions contributed to the correlation for visual detail and effortfulness.

Table 9

Model Fit Statistics for Logistic Mixed Effects Models Predicting Response Accuracy from the Mnemonic Cues in Experiment 2

Fixed effect	$\chi^2(1)$	*p*
Effortfulness	10.46	.001
Visual Detail	11.70	< .001
Clarity	17.40	< .001
Thoughts	5.72	.017
Reasoning	1.83	.176

Table 10

Fixed Effect Coefficients for Logistic Mixed Effects Models Predicting Response Accuracy from the Mnemonic Cues in Experiment 2

Fixed effect	*b*	SE_b	95% CI_b
Effortfulness	-0.45	0.14	-0.73, -0.18
Visual Detail	0.65	0.18	0.29, 1.01
Clarity	0.63	0.13	0.39, 0.88
Thoughts	-0.31	0.13	-0.55, -0.07
Reasoning	-0.16	0.11	-0.38, 0.06

Manipulation checks. To assess the success of my manipulations, I began by examining whether the mnemonic cues and confidence predicted control decisions (i.e., volunteer/withhold decisions). Two logistic mixed effects models were constructed for the clarity + reasoning, clarity + thoughts, and visual detail + effortfulness conditions (one for

each mnemonic cue they were instructed to use), and one model was constructed for the confidence condition. The results showed that model fit was significantly improved with the addition of clarity, thoughts, visual detail, effortfulness, and confidence, while the difference in model fit when reasoning was added approached significance (Table 11). In addition, clarity, thoughts, visual detail, effortfulness, and confidence were all significant predictors of control decisions (Table 12). The confidence interval for reasoning indicated that it bordered on being a significant predictor of control decisions given that the upper limit was 0.00. Thus, the findings were consistent with the notion that participants based their control decisions on the information they were instructed to consider, though perhaps to a lesser extent for reasoning. However, it was also possible that participants were making their control decision before rating the mnemonic cues, and constructing their

Table 11

Model Fit Statistics for Logistic Mixed Effects Models Predicting Control Decisions from the Mnemonic Cues used by Each Condition in Experiment 2

Condition	Mnemonic cue	$\chi^2(1)$	*p*
Confidence	Confidence	54.44	< .001
Clarity + reasoning	Clarity	42.23	< .001
	Reasoning	3.78	.052
Clarity + thoughts	Clarity	59.00	< .001
	Thoughts	12.33	< .001
Visual detail + effortfulness	Visual detail	55.97	< .001
	Effortfulness	17.23	< .001

Table 12

Fixed Effect Coefficients for Logistic Mixed Effects Models Predicting Control Decisions from the Mnemonic Cues Used by Each Condition in Experiment 2

Condition	Mnemonic cue	*b*	SE_b	95% CI_b
Confidence	Confidence	1.78	0.33	1.13, 2.43
Clarity + reasoning	Clarity	3.01	0.49	2.04, 3.97
	Reasoning	-0.95	0.48	-1.89, 0.00
Clarity + thoughts	Clarity	2.74	0.40	1.96, 3.52
	Thoughts	-1.26	0.37	-1.98, -0.55
Visual detail + effortfulness	Visual detail	3.15	0.48	2.21, 4.09
	Effortfulness	-2.63	0.70	-4.05, -1.29

ratings based on their control decision.[28]

I also examined the data for the manipulation check questions to determine whether the instructions changed the participants' conscious awareness and knowledge about the mnemonic cues and confidence.[29] Chi-square tests were used to analyse data from the multiple choice manipulation check questions which asked what information should be considered when judging response accuracy (Table 13) and deciding what information to volunteer (Table 14). A separate analysis was undertaken for each mnemonic cue. The condition(s) that was(were) instructed to use a particular mnemonic cue was(were)

[28] When answering questions, participants were instructed to enter their answer, complete the mnemonic cue ratings, and then make a control decision. However, it is possible that participants made a control decision within their own minds prior to completing their mnemonic cue ratings.

[29] Mixed effects analysis was not required for these data because there was a single observation for each participant for all items.

compared to all other conditions. For example, when analysing clarity, the clarity + reasoning and clarity + thoughts conditions that were instructed to use the clarity cue were compared to the visual detail + effortfulness and confidence conditions. Chi-square tests were used to analyse the data from the questions that asked whether memories should be volunteered or withheld if they were associated with high confidence or an experience of each mnemonic (Table 15). Data for the questions that asked participants to rate how likely memories were to be correct when they were associated with high confidence or an experience of each mnemonic cue were analysed using *t*-tests (Table 16). These ratings were on a five-point scale ranging from 0 (likely to be incorrect) to 4 (likely to be correct).

The results for the manipulation check questions revealed that participant knowledge regarding clarity, visual detail, and confidence was similar regardless of the instructions participants received. The majority of participants knew it was important to consider confidence, visual detail, and clarity when judging response accuracy and making control decisions. Furthermore, regardless of instruction condition, almost all of them knew they should volunteer clear memories, visually detailed memories, and memories associated with high confidence; and a majority of them also knew that clear and visually detailed memories were likely to be correct. Although participants who received confidence instructions rated memories associated with high confidence as being significantly more likely to be correct than participants who did not receive confidence instructions, the mean for both groups was on the *likely to be correct* side of the scale (i.e., above 2). This indicated that most participants were aware that memories associated with high confidence are likely to be correct. These findings suggested that the confidence and visual detail instructions do act as suitable control instructions. However, they also showed that the clarity instructions were unsuccessful in altering knowledge about the

Table 13

Percentage of Participants Citing Confidence, Visual Detail, Effortfulness, Clarity, Reasoning, and Thoughts as Important to Consider when Judging Response Accuracy in Experiment 2

Mnemonic cue	Group 1	Group 2	Group 1 %	Group 2 %	$\chi^2(1)$	*p*	*φ*
Confidence	Confidence	CR, CT & VE	61.90	50.00	0.49	.486	.10
Visual detail	VE	CR, CT & confidence	71.43	42.19	4.30	.038	.25
Effortfulness	VE	CR, CT & confidence	38.10	4.69	12.84	< .001	.43
Clarity	CR and CT	VE & confidence	79.07	52.38	5.60	.018	.28
Reasoning	CR	CT, VE & confidence	31.82	1.59	14.11	< .001	.45
Thoughts	CT	CR, VE & confidence	4.76	9.38	0.04	.834	.20

Note. CR = clarity + reasoning, CT = clarity + thoughts, VE = visual detail + effortfulness. The phi φ coefficient is interpreted in the same way as correlation coefficient. Thus, .10 is considered a small effect .30 is considered a medium effect and .50 is considered a large effect.

Table 14

Percentage of Participants Citing Confidence, Visual Detail, Effortfulness, Clarity, Reasoning, and Thoughts as Important to Consider when Deciding What Information to Volunteer in Experiment 2

Mnemonic cue	Group 1	Group 2	Group1 %	Group 2 %	$\chi^2(1)$	*p*	*φ*
Confidence	Confidence	CR, CT & VE	80.95	51.56	4.49	.034	.26
Visual detail	VE	CR, CT & confidence	42.86	32.81	0.33	.567	.09
Effortfulness	VE	CR, CT & confidence	47.62	14.06	8.42	.004	.35
Clarity	CR and CT	VE & confidence	65.12	34.88	2.64	.104	.20
Reasoning	CR	CT, VE & confidence	50.00	5.35	18.48	< .001	.50
Thoughts	CT	CR, VE & confidence	23.81	6.25	3.46	.063	.25

Note. CR = clarity + reasoning, CT = clarity + thoughts, VE = visual detail + effortfulness. The phi φ coefficient is interpreted in the same way as correlation coefficient. Thus, .10 is considered a small effect .30 is considered a medium effect and .50 is considered a large effect.

Table 15

Percentage of Participants Who Said Memories Should be Volunteered If They Were Associated with High Confidence or an Experience of Each Mnemonic Cue in Experiment 2

	Group 1	Group 2	Group1 %	Group 2 %	$\chi^2(1)$	*p*	*φ*
Confidence	Confidence	CR, CT & VE	100.00	100.00	-	-	-
Visual detail	VE	CR, CT & confidence	96.67	96.67	-	-	-
Effortfulness	VE	CR, CT & confidence	20.00	25.56	0.14	.712	.06
Clarity	CR and CT	VE & confidence	100.00	100.00	-	-	-
Reasoning	CR	CT, VE & confidence	23.22	43.33	3.01	.083	.18
Thoughts	CT	CR, VE & confidence	16.67	25.56	0.56	.455	.09

Note. CR = clarity + reasoning, CT = clarity + thoughts, VE = visual detail + effortfulness. The phi φ coefficient is interpreted in the same way as correlation coefficient. Thus, .10 is considered a small effect .30 is considered a medium effect and .50 is considered a large effect.

Table 16

Means and Standard Deviations for How Likely Participants Thought Memories Were to be Correct When They Were Associated with High Confidence or an Experience of Each Mnemonic Cue in Experiment 2

Mnemonic cue	Group 1	Group 2	Group1 mean	Group 2 mean	*t*	*df*	*p*	Cohen's *d*
Confidence	Confidence	CR, CT & VE	3.40 (0.67)	2.67 (1.11)	4.07	93.13	< .001	0.73
Visual detail	VE	CR, CT & confidence	3.00 (1.31)	3.24 (0.88)	0.95	38.02	.348	-0.25
Effortfulness	VE	CR, CT & confidence	1.10 (1.03)	2.02 (1.02)	4.26	49.24	< .001	0.91
Clarity	CR and CT	VE & confidence	3.28 (0.96)	3.20 (1.10)	0.44	115.78	.659	0.08
Reasoning	CR	CT, VE & confidence	1.63 (1.30)	2.16 (1.02)	2.01	41.46	.051	0.48
Thoughts	CT	CR, VE & confidence	1.43 (1.14)	1.95 (1.08)	2.06	47.85	.044	0.48

Note. CR = clarity + reasoning, CT = clarity + thoughts, VE = visual detail + effortfulness. *SD*'s in parentheses. Scores can range from 0 (likely to be incorrect) to 4 (likely to be correct).

clarity cue, and instead suggested that witnesses may base their monitoring decisions on confidence, visual detail, and clarity without being encouraged to do so.

The manipulation check question results also revealed that the thoughts instructions had only a minimal impact on knowledge about the thoughts cue. The findings showed that very few participants said it was important to consider thoughts when judging response accuracy and making control decisions, regardless of the instructions they received. There was also no significant difference in the percentage of participants who said memories including a lot of thoughts should be volunteered, though there seemed to be a consensus that memories that include a lot of thoughts should not be volunteered. The only indication that the thoughts instructions had any impact of participant knowledge about this cue was that participants who received thoughts instructions rated memories that include a lot of thoughts as being significantly more likely to be incorrect than participants who did not receive thoughts instructions. Together, these results showed that the manipulation of the thoughts cue was generally unsuccessful.

Although it appeared that the manipulations of the clarity and thoughts cues were unsuccessful, the reasoning and effortfulness instructions did appear to have altered participant knowledge regarding reasoning and effortfulness, respectively. A higher percentage of participants who received reasoning instructions said reasoning was important to consider when judging response accuracy and making control decisions compared to participants who did not receive reasoning instructions. Participants who received reasoning instructions also rated memories involving reasoning as more likely to be incorrect than participants who did not receive reasoning instructions. Although this difference did not quite reach significance, the effect size was moderate, suggesting that the difference was meaningful. Furthermore, the mean for participants who received reasoning instructions was on the *likely to be incorrect* side of the scale (i.e., above 2),

while the mean for participants who did not receive reasoning instructions sat roughly in the centre of the scale. Regarding effortfulness, s higher percentage of participants who received effortfulness instructions said that effortfulness was important to consider when judging response accuracy and making control decisions compared to participants who did not receive effortfulness instructions. Participants who received effortfulness instructions also rated memories perceived as being effortful to retrieve as significantly more likely to be incorrect than participants who did not receive effortfulness instructions. The fact that both reasoning and effortfulness instructions altered the participants' knowledge of these mnemonic cues suggests that people may not consider them to be as important as they should and that they may not use them to their full potential during monitoring. .

Overall, it appeared that the instructions provided increased awareness and knowledge of the reasoning and effortfulness cues, but not the clarity, thoughts, or visual detail cues. The clarity, thoughts, and reasoning instructions were expected to increase awareness and knowledge of the clarity, thoughts, and reasoning cues, respectively, because Study 1 suggested that these mnemonic cues are not fully utilised during monitoring. However, the visual detail and effortfulness instructions were not expected to increase awareness and knowledge of the visual detail and effortfulness cues, respectively, because Study 1 suggested that witnesses do fully utilise these mnemonic cues during monitoring. The fact that the clarity and visual detail instructions did not have an impact on awareness and knowledge of clarity and visual detail, respectively, coupled with the fact that most participants seemed to be aware of the importance and utility of these cues, suggested that clarity and visual detail may be considered to their full extent during monitoring. As very few participants seemed aware of the importance and utility of the thoughts cue, regardless of the instructions that were provided, it appears that this manipulation was unsuccessful. However, the fact that the reasoning and effortfulness

instructions did have an impact on awareness and knowledge of reasoning and effortfulness, respectively, suggested that these cues may not be fully utilised during monitoring.

The effect of instruction type on monitoring and response bias. To assess the impact of instruction type on monitoring (i.e., type-2 discriminability), I constructed a logistic mixed effects model which compared the clarity + reasoning, clarity + thoughts, and visual detail + effortfulness conditions to the control condition. This analysis also allowed me to assess the impact of the instructions on response bias. The results of the analysis are presented in Figure 2 and the coefficients are presented in Table 17. As explained in Chapter 2, monitoring is represented by the relative difference between the bars for correct and incorrect responses, while response bias is represented in the figure by the combined height of the correct and incorrect bars.

The results showed that monitoring did not differ significantly between the confidence condition and any of the experimental conditions. Figure 2 shows the relative difference between the bars for correct and incorrect responses is similar across conditions. However, in comparison to the confidence condition, the difference is slightly smaller for the clarity + reasoning and clarity + thoughts conditions and slightly larger for the visual detail + effortfulness condition. Consistent with this, the coefficients for the clarity + reasoning and clarity + thoughts condition were negative, indicating that these conditions had less effective monitoring than the confidence condition, though not significantly so. In addition, the positive coefficient for the visual detail + effortfulness condition indicated that this condition had more effective monitoring than the confidence condition, though not significantly so. The lack of a significant difference in monitoring was confirmed by the model fit comparison which showed that adding the interaction between condition and response accuracy to the model did not significantly improve fit, χ^2 (3) = 1.51, p = .680.

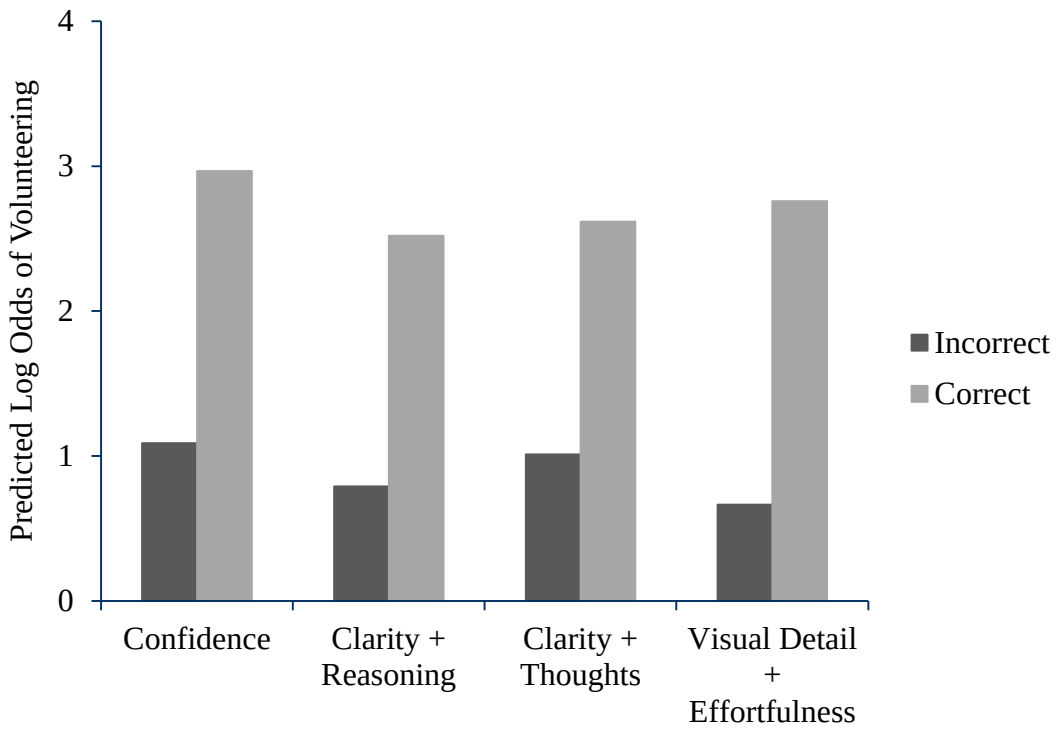

Figure 2. Predicted log odds of volunteering correct and incorrect answers as a function of the type of instructions provided in Experiment 2.

The results also showed that response bias did not differ significantly between the confidence condition and any of the experimental conditions. Figure 2 also shows that the combined height of the correct and incorrect bars is similar across conditions; though it is slightly lower for each experimental condition than for the confidence condition. Indeed, all the coefficients for response bias were negative. This indicated that the experimental conditions were less likely to volunteer than the confidence condition, though not significantly so. The lack of a significant difference in response bias was confirmed by the model fit comparison which showed that adding condition to the model did not significantly improve fit, χ^2 (3) = 1.81, p = .613. In addition, there was a very strong bias against withholding as the mean proportion of withheld responses was just 0.16 (*SD* =

0.14) across conditions while the mean proportion of incorrect responses was .33 (*SD* = 0.12).

Table 17

Fixed Effects Coefficients for Response Bias and Monitoring for the Clarity + Reasoning, Clarity + Thoughts, and Visual Detail + Effortfulness Conditions in Comparison to the Confidence Condition in Experiment 2

	Response bias		
Condition	*b*	SE_b	95% CI_b
Clarity + reasoning	-0.41	0.34	-1.07, 0.26
Clarity + thoughts	-0.39	0.34	-1.05, 0.28
Visual detail + effortfulness	-0.35	0.34	-1.02, 0.33
	Monitoring		
Clarity + reasoning	-0.15	0.40	-0.93, 0.64
Clarity + thoughts	0.22	0.41	-0.59, 1.02
Visual detail + effortfulness	-0.27	0.41	-1.08, 0.54

The effect of instruction type on accuracy and quantity. As there was no evidence that monitoring was significantly improved (or impaired) as a result of the instructions, I did not expect any significant differences in accuracy or quantity between the conditions, despite my initial predictions. Consistent with this, the results of a logistic mixed effect analysis showed that accuracy did not differ significantly between the control condition and any of the experimental conditions (see Table 18 for coefficients and Table 19 for estimated proportions). The lack of a significant difference in accuracy was

confirmed by the model fit comparison which showed that adding condition to the model did not significantly improve fit, χ^2 (3) = 0.88, p = .830. Similarly, the results of another logistic mixed effects model showed that quantity did not differ significantly between the control condition and any of the experimental conditions. The lack of a significant difference in quantity was confirmed by the model fit comparison which showed that adding condition to the model did not significantly improve fit, χ^2 (3) = 2.74, p = .434.

Table 18

Fixed Effects Coefficients for Accuracy and Quantity for the Clarity + Reasoning, Clarity + Thoughts, and Visual Detail + Effortfulness Conditions in Comparison to the Confidence Condition in Experiment 2

	Accuracy		
Condition	*b*	SE_b	95% CI_b
Clarity + reasoning	0.10	0.31	-0.50, 0.70
Clarity + thoughts	-0.14	0.25	-0.63, 0.35
Visual detail + effortfulness	-0.19	0.26	-0.69, 0.31
	Quantity		
Clarity + reasoning	-0.23	0.24	-0.69, 0.23
Clarity + thoughts	-0.33	0.24	-0.79, 0.14
Visual detail + effortfulness	-0.36	0.24	-0.82, 0.11

Table 19

Estimated Proportions for Accuracy and Quantity for Each Condition in Experiment 2

Condition	Accuracy	Quantity
Confidence	.75	.73
Clarity + reasoning	.76	.68
Clarity + thoughts	.72	.65
Visual detail + effortfulness	.71	.65

Discussion

My aim in Experiment 2 was to test whether mnemonic cue instructions could improve monitoring and subsequently allow people to maximise the quantity and accuracy of their eyewitness memory reports. Contrary to predictions, the results revealed that monitoring was not improved by providing witnesses with information about the clarity cue in combination with either the reasoning cue or the thoughts cue. In line with this, the mnemonic cue instructions did not have a significant impact on quantity or accuracy. The lack of improvement in monitoring for the clarity + thoughts condition was likely due the unsuccessful manipulation, and it was possible the clarity + reasoning instructions were unable to improve monitoring because reasoning is not a reliable predictor of response accuracy. I would not expect providing information about a mnemonic cue that is not diagnostic of response accuracy to improve monitoring. However, there are two other possible explanations regarding the ineffectiveness of the clarity + reasoning instructions which are discussed in detail below. These explanations could also explain why the visual detail + effortfulness instructions did not have an impact on monitoring, quantity, or accuracy.

It was possible that the mnemonic cue instructions did not improve monitoring because of the way participants made their control decisions. Overall, participants withheld a mean proportion of 0.16 responses (*SD* = 0.14). This was despite the fact that they retrieved a mean proportion of 0.33 incorrect responses (*SD* = 0.12). While this could indicate that the participants were simply overestimating the accuracy of their memory, it was also possible that the response format of the questions biased them towards volunteering. Several studies have explored similar commitment effects for mugshot exposure prior to a line-up (Brigham & Cairns, 1988; Dysart, Lindsay, Hammond, & Dupuis, 2001; Goodsell, Gronlund, & Neuschatz, 2015; Gorenstein & Ellsworth, 1980; Memon, Hope, Barlett, & Bull, 2002). Research employing the mugshot commitment design presents participants with a target face and has them view a series of mugshots that do not include the offender. Participants are encouraged to select one of the innocent individuals from the mugshots. During a subsequent line-up, participants are highly likely to falsely identify the person from the mugshot as the offender, even when the actual offender is present. Although no theoretical account if offered for this commitment effect in the literature, it is possible that the participants from Experiment 2 were also experiencing a commitments effect. Specifically, as participants in Experiment 2 entered each answer before making a control decision, they may have felt committed to the answers they entered, and consequently felt compelled to volunteer them, though this explanation is purely speculative. The instructions given to participants may not have been strong enough to overcome this bias towards volunteering.

It was also possible that the reasoning and effortfulness instructions were unable to improve monitoring because they were overshadowed by the clarity and visual detail instructions. The results for the manipulation check questions indicated that participants who received instructions about clarity and visual detail did not differ from the other

participants in terms of their level of knowledge about these cues. However, the results also showed that the reasoning and effortfulness instructions did impact on participants' knowledge about these cues. It may be that when participants were given instructions about clarity and reasoning or visual detail and effortfulness, they focussed on using the mnemonic cue that was familiar to them (i.e., clarity or visual detail) at the expense of using the mnemonic cue that could have been more helpful (i.e., reasoning or effortfulness). Thus, the reasoning and effortfulness instructions may have been ineffective because, although they increased knowledge about an appropriate mnemonic cue, they may not have changed usage of the appropriate mnemonic cue. In line with this explanation, the results showed that reasoning bordered on being a significant predictor of control decisions. Conversely, effortfulness was found to be a significant predictor of control decisions. However, it may be that effortfulness was not being used extensively enough to have an effect on monitoring. Indeed, the coefficients indicated that visual detail was a stronger predictor of control decisions than effortfulness.

In sum, Experiment 2 provided evidence that mnemonic cue instructions do not improve monitoring or allow people to maximise the quantity and accuracy of their eyewitness memory reports. However, it was also possible that the mnemonic cue instructions were unable to improve monitoring because the method via which participants responded to questions produced an extremely strong bias towards volunteering. Furthermore, it could also be that improvements in monitoring were prevented by the mnemonic cue combinations that were used because the mnemonic cue pairs consisted of a familiar and regularly used cue and a potentially less familiar and used cue. These possibilities were explored in Experiment 3 by altering the mnemonic cue instructions and having participants answer the recall questions in two phases; an initial free-report phase

and second phase in which they provided responses to any unanswered questions from phase one.

Experiment 3

Experiment 3 aimed to further explore whether providing people with information about particular mnemonic cues can improve their monitoring ability and allow them to maximise the quantity and accuracy of their eyewitness memory reports. The basic procedure of Experiment 3 was the same as Experiment 2. However, I altered the way participants made their control decisions based on the findings obtained in Experiment 2. In the discussion section of Experiment 2, I explained that the response format of the questions may have led participants to avoid withholding. Specifically, participants could have felt compelled to volunteer their responses to the recall questions because they entered them prior to making a control decision. Therefore, to address this limitation, Experiment 3 utilised a two-phase reporting procedure which more closed aligned with the approaches used in the literature for obtaining responses (see Koriat & Goldsmith, 1996; Weber & Brewer, 2008). Participants first answered the recall questions in a free-report manner by either volunteering an answer or saying *don't know*. In a second forced-report phase, participants provided responses to all the questions for which they selected the *don't know* response, as this data is required in order to conduct the Type-2 SDT analyses. While the phases are typically presented in the opposite order in the literature (i.e., forced-report followed by free-report), presenting the free-report phase first meant that the experiment duration could be minimised as not all questions needed to be repeated in the forced-report phase.

Experiment 3 consisted of one control condition and two experimental conditions. In contrast to Experiment 2, the control condition did not receive any cue-related instructions or instructions about confidence. The decision not to include a no-instruction

control condition in Experiment 2 was based on Koriat and Goldsmith's (1996) metamemory framework which proposes that people use confidence to judge response accuracy. However, in retrospect, I acknowledged the possibility that while people may use confidence during monitoring without being encouraged to do so, explicitly instructing them to consider confidence may cause them to examine their memories more closely than they would without such instruction, potentially leading to improved monitoring. Thus, a no-instruction control condition was included in Experiment 3. The first experimental condition received reasoning instructions while the second received effortfulness instructions. Instructions about clarity and visual detail were omitted for two reasons. First, the results of Experiment 2 indicated that people consider these cues during monitoring without being instructed to,[30] and therefore, instructions about them would not be expected to improve monitoring. Second, it was possible that the inclusion of clarity and visual detail instructions prevented the reasoning and effortfulness instructions from having an impact on monitoring in Experiment 2, as explained in the discussion section for that experiment. Thus, omitting the clarity and visual detail instructions in Experiment 3 allowed me to examine this as a potential explanation for the findings observed in Experiment 2.

Including conditions that only received reasoning or effortfulness instructions, also allowed me to explore the inconsistent findings of Study 1 and Experiment 2. Although reasoning was found to be a significant predictor of response accuracy in Study 1, it was not in Experiment 2, and it was not possible to determine whether the result for Study 1 was a Type I error or whether the result for Experiment 2 was a Type II error. Similarly, the results of Study 1 suggested that people may consider effortfulness during monitoring without being instructed to, while the findings of Experiment 2 suggested the opposite and

[30] Study 1 also suggested that people consider visual detail during monitoring without instruction.

indicated that effortfulness may be a reliable predictor of response accuracy. Therefore, including experimental conditions that either received reasoning or effortfulness instructions (and ratings) allowed me to obtain further evidence about how these cues relate to accuracy and whether encouraging their use can improve monitoring.

I predicted that the reasoning and effortfulness conditions would have better monitoring (i.e., type-2 discriminability) than the control condition. However, as explained in the introduction of Experiment 2, the impact of any improvements in monitoring on quantity and accuracy could manifest in different ways. For accuracy to be improved there needs to be (i) an increase in the amount of correct information being volunteered, (ii) a decrease in the amount of incorrect information being volunteered, or (iii) a combination of both. Each of these options constitutes an improvement in monitoring. Thus, if the reasoning and effortfulness conditions exhibited better monitoring than the control condition, accuracy was also expected to be higher for the reasoning and effortfulness conditions than for the control conditions. However, to avoid reducing quantity, the proportion of retrieved information that is both correct and volunteered must be maximised. This means that I only expected quantity to be higher for the reasoning and effortfulness conditions if they volunteered a higher amount of correct information than the control condition.

Method

Participants. Ninety participants[31] (66 females and 24 males) took part in the study for course credit or payment ($15). All had normal or corrected-to-normal vision and spoke English as their first language. Participants ranged in age from 17 to 39 (*M* = 21.84, *SD* = 5.13).

[31] Before data collection began we decided to recruit 90 participants, 30 per condition. I ended data collection in the week that I reached 90 participants.

Materials.

Stimulus video and recall task. Participants viewed the same video that was used in Study 1 and Experiment 2 and completed the same modified recall task from Experiment 2.

Monitoring instructions. Participants either received reasoning instructions, effortfulness instructions or no cue-related or confidence instructions. The reasoning instructions informed participants that memories involving reasoning are likely to be incorrect, and that they should respond *don't know* when their memory involves reasoning. Similarly, the effortfulness instructions informed participants that memories that are effortful to retrieve are likely to be incorrect, and that they should respond *don't know* when their memory is effortful to retrieve.

Mnemonic cue rating scales. Participants in the reasoning and effortfulness conditions completed ratings of reasoning and effortfulness, respectively, in the same way as Study 1 and Experiment 2. As explained in the methodology section of Experiment 2, these ratings constituted a second part of the manipulation as they served as a reminder to the participants that they should consider the mnemonic cues throughout the recall task.

Manipulation check questions. The first two manipulation check questions were open-ended and asked what information participants should use when judging whether their memory is correct and what information they should use when deciding whether to volunteer an answer or respond *don't know*. These questions were included to determine whether participants could spontaneously recall the information given in the instructions. A marking guide was constructed to code these responses, and coding was completed by me and an independent rater on a sample of the data from 12 randomly selected participants. There was an acceptable level of interrater agreement, $\kappa = .81$ [.68, .94].

Responses were coded using seven categories: reasoning, effortfulness, confidence, visual detail, retrieval fluency, don't know, and other. Some participants were assigned several codes as they provided multiple responses.

Participants also completed 10 other manipulation check questions that asked how important different types of information were to consider when judging correctness and making *don't know* decisions. Although I was primarily interested in the participants' perceptions of the importance of considering effortfulness and reasoning, I also included questions about retrieval fluency, confidence, and visual detail to minimise demand effects. Participants rated the importance of each cue on a five-point rating scale ranging from 0 (not at all important) to 4 (extremely important).

Procedure. After giving informed consent, participants were randomly assigned to one of three levels of the only manipulation: reasoning, effortfulness, or control (n = 30 in each condition). Participants were shown the video and completed the 10-min maze filler task that was used in Study 1 and Experiment 2. All participants completed the experiment on a computer.

After the filler task, participants received their monitoring instructions (the exact wording of all instructions is provided in Appendix D). These instructions were repeated after a practice question (also provided in Appendix D), and participants then completed the recall task. When presented with filter questions, participants had two response options, *yes* and *no*. When they said *yes*, they were presented with a closed follow-up question. Participants were asked to think of their best possible answer before proceeding. After indicating that they had retrieved an answer, but without reporting the answer, participants in the reasoning and effortfulness conditions rated reasoning and effortfulness, respectively. Participants in the control condition did not complete any ratings. All participants were then asked whether they would like to provide the answer they had in

mind or say *don't know*. When they indicated that they wanted to provide the answer, the question was presented again with a text box to type the answer. When participants had responded to all of the questions, they were asked to provide an answer to the questions that had received a *don't know* response. Participants were asked to think back to the answer they had in mind when they were first presented with the question. Before being presented with these questions, participants were told that although saying *don't know* when they are unsure is good; the purpose of asking for responses to these questions was to get an idea of what their best guess was. No other ratings were made in this phase of the experiment. Upon completion of the questions, all participants responded to the manipulation check questions.

For each question in the recall task I measured only: (i) the participant's response, (ii) the score they gave on the rating scale (reasoning and effortfulness conditions only), (iii) whether they chose to volunteer or withhold the answer (closed questions only), and (iv) the reaction time for each of these responses. Responses to the manipulation check questions, and the reaction times of these responses were also recorded (none of the reaction time data were analysed).

Results

The relationship between mnemonic cues and response accuracy. As in Experiment 2, I assessed whether the mnemonic cues predicted response accuracy. The results of a logistic mixed effects model showed that reasoning was not a significant predictor of response accuracy, b = -0.10, SE_b = 0.09, [-0.28, 0.08], and adding it to the model did not significantly improve fit, χ^2 (1) = 1.20, p = .272.[32] However, the results of another logistic mixed effects model showed that effortfulness was a significant predictor

[32] As only participants in the reasoning condition completed these ratings, this analysis only includes the reasoning condition.

of response accuracy, b = -0.26, SE_b = 0.09, [-0.43, -0.08], and its addition to the model significantly improved fit, χ^2 (1) = 8.04, p = .004.[33] Thus, it appeared that while the effortfulness cue was a reliable predictor of response accuracy, the reasoning cue was not.

Manipulation checks. To assess the success of the manipulations, I initially examined whether reasoning and effortfulness predicted control decisions. The results of a logistic mixed effect model showed that reasoning was a significant predictor of control decisions, b = -4.06, SE_b = 1.11, [-6.23, -1.89], and adding it to the model significantly improved fit, χ^2 (5) = 312.46, p < .001. Similarly, effortfulness was a significant predictor of control decisions, b = -2.89, SE_b = 0.84, [-4.53, -1.25], and adding it to the model significantly improved fit, χ^2 (5) = 332.75, p < .001. The direction of these relationships was consistent with the instructions participants received. Specifically, participants in the reasoning and effortfulness conditions were more likely to withhold responses that were associated with high reasoning or effortfulness ratings, respectively. Thus, the findings were consistent with the notion that participants based their control decisions on the information they were instructed to consider. However, it is also possible that participants were making their control decision before rating the mnemonic cue, and constructing their rating based on their control decision.[34]

Next, I examined data for the open-ended manipulation check questions that asked what information should be considered when judging response accuracy and responding *don't know*. I used chi-square tests to assess the percentage of participants who cited

[33] As only participants in the effortfulness condition completed these ratings, this analysis only includes the effortfulness condition.

[34] When answering questions, participants were instructed to bring an answer to mind, complete the mnemonic cue rating, and then make a control decision. However, it is possible that participants made a control decision within their own minds prior to completing their mnemonic cue rating.

reasoning and effortfulness in response to these questions (Table 20).[35] The reasoning condition was compared to a group containing the effortfulness and control conditions, and the effortfulness condition was compared to a group containing the reasoning and control conditions. The results showed that a higher percentage of participants in the reasoning condition cited reasoning as important to consider when judging response accuracy and responding *don't know* compared to the effortfulness and control conditions. Similarly, a higher percentage of participants in the effortfulness condition cited effortfulness as important to consider when judging response accuracy and responding *don't know* compared to the reasoning and control conditions. Thus, it appeared that the instructions did alter the participants' perceptions of what information is important to consider when judging response accuracy and responding *don't know*.

However, the percentages also indicated that the manipulation may have been ineffective for a large number of participants. In the effortfulness condition the percentage of participants citing effortfulness was quite low (i.e., below 30%), and in the reasoning condition the percentage of participants citing reasoning did not exceed 50%. These percentages suggested that many participants either did not remember the instructions, or were unable to spontaneously recall them, which could indicate that the manipulations were unsuccessful. Alternatively, it was possible that participants did not realise they were being asked about the instructions they received. For each open-ended manipulation check question, 14.44% of participants either left the question blank, or said they were unsure about what information they should consider when judging response accuracy and responding *don't know*. Thus, the participants may have been confused about how to answer the open-ended manipulation check questions. Some may have thought the

[35] Mixed effects analysis was not required for these data because there was a single observation for each participant for all items.

Table 20

Percentage of Participants Who Said Reasoning and Effortfulness Were Important to Consider When Judging Response Accuracy and Responding Don't Know in Experiment 3

	Judging accuracy						
Mnemonic cue cited	Group 1	Group 2	Group 1 %	Group 2 %	$\chi^2(1)$	*p*	*φ*
Reasoning	R	E & C	33.33	5.00	10.80	.001	.38
Effortfulness	E	R & C	26.67	3.33	8.79	.003	.35
	Responding don't know						
Reasoning	R	E & C	43.33	6.67	15.24	< .001	.44
Effortfulness	E	R & C	16.67	0.00	7.65	.005	.34

Note. R = reasoning, E = effortfulness, C = control

question was asking what information they thought was important to consider, others may have thought that it was asking what information they had been told was important to consider, and some obviously did not understand what was being asked. Thus, it was difficult to judge the effectiveness of the mnemonic cue instructions based on the results for the open-ended manipulation check questions. However, based on the assumption that these manipulation check questions were assessing the information participants were told was important, it appeared that the reasoning manipulation may have only been successful for approximately half of participants, while the effortfulness manipulation may have only been successful for approximately a third of participants.

Finally, I examined the manipulation check questions that asked how important it is to consider reasoning and effortfulness when judging response accuracy and responding *don't know*. The data were analysed using *t*-tests and a separate analysis was conducted for each question (Table 21).[36] For questions that asked about reasoning, the reasoning condition was compared to a group containing the effortfulness and control conditions, while for questions that asked about effortfulness, the effortfulness condition was compared to a group containing the reasoning and control conditions. The results showed that participants in the reasoning condition rated reasoning as significantly more important to consider when making *don't know* decisions compared to the other conditions, though they did not rate reasoning as significantly more important to consider when judging response accuracy. However, the latter difference did approach significance and the effect size was moderate which suggested that the difference may be meaningful. Participants in the effortfulness condition rated effortfulness as significantly more important to consider when judging response accuracy compared to participants in the reasoning and control

[36] Mixed effects analysis was not required for these data because there was a single observation for each participant for all items.

Table 21

Means and Standard Deviations for How Important Participants Thought Reasoning and Effortfulness Were To Consider When Judging Response Accuracy and Responding Don't Know in Experiment 3

	Judging correctness							
Mnemonic cue	Group 1	Group 2	Group1 mean	Group 2 mean	*t*	*df*	*p*	Cohen's *d*
Reasoning	R	E & C	2.93 (1.05)	2.52 (0.81)	1.91	46.97	.062	0.47
Effortfulness	E	R & C	3.17 (0.70)	2.47 (0.99)	4.02	76.43	< .001	0.77
	Responding don't know							
Reasoning	R	E & C	2.90 (0.92)	0.96 (0.57)	2.55	59.99	.013	0.57
Effortfulness	E	R & C	2.77 (0.94)	2.56 (0.99)	1.03	59.72	.309	0.22

Note. R = reasoning, E = effortfulness, C = control. *SD*'s in parenthesis. Scores can range from 0 (not at all important) to 4 (extremely important).

conditions, though they did not rate effortfulness as significantly more important to consider when making *don't know* decisions. This could indicate that participants were not basing their control decisions on effortfulness, though the mixed effects analysis suggested that effortfulness was predictive of control decisions.

Overall, it appeared that the reasoning instructions were successful in changing the participants' awareness and knowledge about reasoning, though it is possible that the manipulation was not successful for all participants based on the open-ended manipulation check questions. The success of this manipulation may be problematic given that reasoning was not found to be a significant predictor of response accuracy. Encouraging usage of a mnemonic cue that is not predictive of response accuracy would not be expected to improve monitoring and could even worsen monitoring if its usage results in the neglect of other mnemonic cues that are diagnostic of response accuracy. The results also indicated that the effortfulness manipulation may have only been successful for around a third of participants.

The effect of instruction type on monitoring and response bias. To assess the impact of instruction type monitoring (i.e., type-2 discriminability), a logistic mixed effects model was constructed that compared the reasoning and effortfulness conditions to the control condition. This analysis also allowed me to assess the impact of the instructions on response bias. The results of the analysis are presented in Figure 3 and the coefficients are presented in Table 22. Recall that monitoring is represented by the relative difference between the bars for correct and incorrect responses and response bias is represented by the combined height of the correct and incorrect bars.

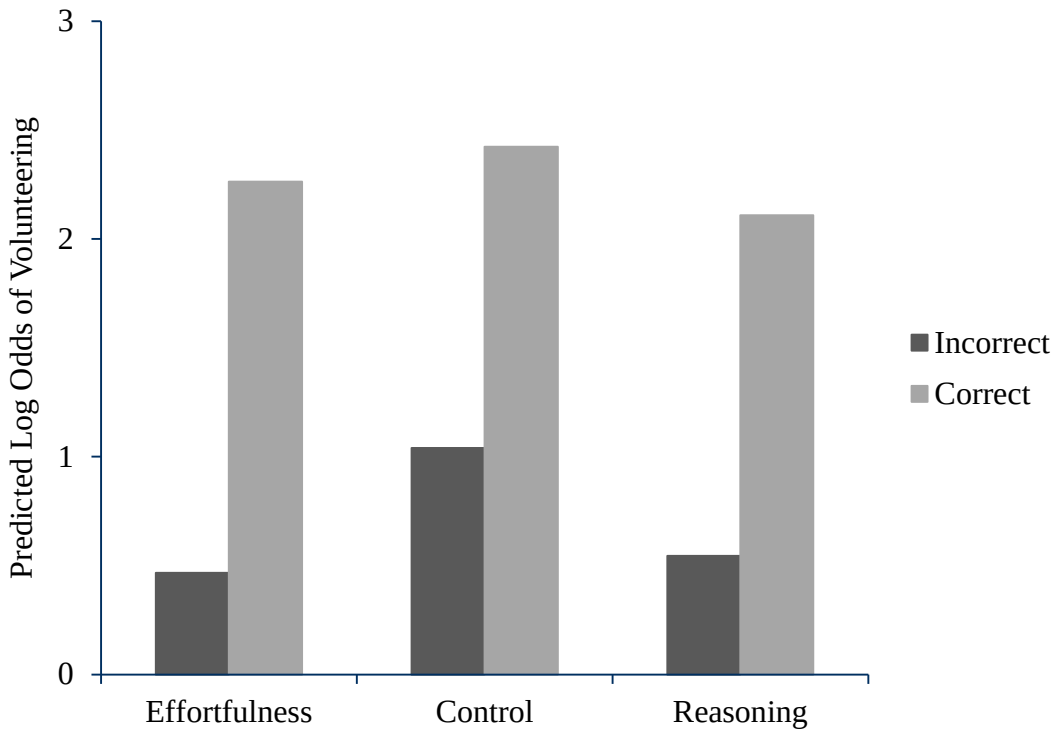

Figure 3. Predicted log odds of volunteering correct and incorrect answers as a function of the type of instructions provided in Experiment 3.

Table 22

Fixed Effects Coefficients for Response Bias and Monitoring for the Reasoning and Effortfulness Conditions in Comparison to the Control Condition in Experiment 3

	Response bias		
Condition	*b*	SE_b	95% CI_b
Reasoning	-0.34	0.33	-0.98, 0.30
Effortfulness	-0.25	0.33	-0.89, 0.40
	Monitoring		
Reasoning	0.41	0.43	-0.40, 1.23
Effortfulness	0.18	0.40	-0.60, 0.97

The results showed that monitoring did not differ significantly between the control condition and the reasoning or effortfulness conditions. Figure 3 shows that the relative difference between the bars for correct and incorrect responses is similar across conditions. However, the difference is slightly larger for the reasoning and effortfulness conditions than for the control condition. Consistent with this, positive coefficients were observed for the reasoning and effortfulness conditions, indicating that these conditions had more effective monitoring than the control condition, though not significantly so. The lack of a significant difference in monitoring was confirmed by the model fit comparison which showed that adding the interaction between condition and response accuracy to the model did not significantly improve fit, χ^2 (2) = 0.93, p = .627.

The results also showed that response bias did not differ significantly between the control condition and the reasoning or effortfulness conditions. Figure 3 shows that the combined height of the correct and incorrect bars is similar across conditions; though it is slightly lower for the reasoning and effortfulness conditions than for the control condition. Indeed, both coefficients for response bias were negative. This indicated that the reasoning and effortfulness conditions were less likely to volunteer than the control condition, though not significantly so. The lack of a significant difference in response bias was confirmed by the model fit comparison which showed that adding condition to the model did not significantly improve fit, χ^2 (2) = 1.11, p = .574. In addition, there was a very strong bias against withholding as the mean proportion of withheld responses was just 0.19 (*SD* = 0.13) across conditions.

The effect of instruction type on accuracy and quantity. As there was no evidence that monitoring was significantly improved (or impaired) as a result of the instructions, I did not expect any significant differences in accuracy or quantity between the conditions, despite my initial predictions. Consistent with this, the results of a logistic

mixed effects model showed that accuracy did not differ significantly between the control condition and the reasoning or effortfulness conditions (see Table 23 for coefficients and Table 24 for estimated proportions). The lack of a significant difference in accuracy was confirmed by the model fit comparison which showed that adding condition to the model did not significantly improve fit, χ^2 (2) = 1.61, p= .448. Similarly, the results for another logistic mixed effects model showed that quantity did not differ significantly between the

Table 23

Fixed Effect Coefficients for Logistic Mixed Effects Models Predicting Accuracy and Quantity from Instructions Type in Experiment 3

	Accuracy		
Instruction type	*b*	SE_b	95% CI_b
Reasoning	0.33	0.29	-0.25, 0.91
Effortfulness	0.29	0.26	-0.22, 0.79
	Quantity		
Reasoning	0.07	0.26	-0.44, 0.59
Effortfulness	-0.03	0.26	-0.53, 0.48

Table 24

Estimated Proportions for Accuracy and Quantity for Each Condition in Experiment 3

Condition	Accuracy	Quantity
Control	.73	.50
Reasoning	.79	.52
Effortfulness	.78	.50

control condition and reasoning or effortfulness conditions. The lack of a significant difference in quantity was confirmed by the model fit comparison which showed that adding condition to the model did not significantly improve fit, χ^2 (2) = 0.14, p = .930.

Discussion

The aim of Experiment 3 was to further explore whether mnemonic cue instructions could improve monitoring and subsequently allow people to maximise the quantity and accuracy of their eyewitness memory reports. The results showed that reasoning and effortfulness instructions did not produce an improvement in monitoring, and quantity and accuracy were not significantly affected by the instructions. It is likely that the reasoning instructions were unable to improve performance because this cue is not a reliable predictor of response accuracy. I will return to this point in the general discussion of this chapter as it is also relevant to Experiment 2. There was also some indication that the effortfulness instructions may not have changed the way participants used this cue when making control decisions, which could explain why receiving instructions about effortfulness did not improve performance. This finding could indicate that people consider effortfulness when making control decisions without being encouraged to do so. However, if this were the case, I would not have expected participants who received effortfulness instructions to rate effortfulness as significantly more important to consider when judging response accuracy than participants who did not receive such instructions. Thus, two other explanations for the ineffectiveness of the effortfulness instructions are more plausible. First, participants may not have remembered or believed the effortfulness instructions, meaning they did not use effortfulness to make their control decisions. Second, the effortfulness instructions may not have been strong enough to affect a change in control decisions.

The findings of Experiment 3 also allowed me to rule out the explanations I proposed for the findings observed in Experiment 2. The fact that monitoring, quantity, and accuracy remained unchanged by the reasoning and effortfulness instructions even though the clarity and visual detail instructions were omitted in Experiment 3 suggested that the clarity and visual detail cues were not overshadowing the reasoning and effortfulness cues in Experiment 2. Similarly, the inability of the mnemonic cue instructions to improve monitoring in Experiment 2 cannot be explained by the response format that was used, which asked participants to enter their responses prior to making a control decision. If response format was responsible for the low withholding rate observed in Experiment 2, I would have expected higher withholding rates in Experiment 3 because participants did not know they would be asked to answer the questions to which they responded *don't know*. Contrary to this, similarly low withholding rates were observed across both studies, suggesting that the low withholding rates observed in Experiment 2 were not the result of a commitment effect.

General Discussion

The results of Experiments 2 and 3 suggested that providing people with information about mnemonic cues does not improve their monitoring ability or help them maximise the quantity and accuracy of their eyewitness memory reports. In contrast to Study 1, reasoning ratings did not reliably predict response accuracy, which is likely to explain why the reasoning instructions were unsuccessful. However, this explanation did not hold for the effortfulness instructions as ratings of effortfulness were found to reliably predict response accuracy across Experiments 2 and 3, though these ratings did not predict accuracy in Study 1. A brief discussion of the discrepancies between Study 1 and Experiments 2 and 3 is provided below. Following this discussion, a potential explanation for the inability of the effortfulness instructions to improve monitoring is considered.

Finally, a potential reason behind the low withholding rates observed across Experiments 2 and 3 is explored.

The discrepancies observed between Study 1 and Experiments 2 and 3 may be explained by differences in the cues participants were considering across these studies. In Study 1, participants completed ratings for seven mnemonic cues while in Experiments 2 and 3 participants only completed ratings for one or two mnemonic cues. It is possible that reasoning was found to be a significant predictor of response accuracy in Study 1 because this rating was informed by some of the other ratings. For example, reasoning was highly correlated with both thoughts and effortfulness in Study 1. Thus, participants may have been using effortfulness and thoughts to make their reasoning judgments, and when these mnemonic cues were no longer considered alongside reasoning in Experiments 2 and 3, participants could not adequately judge reasoning. Similarly, it is possible that rating reasoning and thoughts detracted from the effortfulness ratings in Study 1. For example, it may be that much of what is considered to be effortfulness was captured in the reasoning and thoughts ratings in Study 1, reducing the ability of the effortfulness rating to predict response accuracy. Thus, in Experiments 2 and 3, when reasoning and thoughts were not rated in conjunction with effortfulness, participants may have considered these cues when making their effortfulness ratings, causing the effortfulness ratings to be a stronger predictor of response accuracy.

The effortfulness instructions may have been unable to improve monitoring because witnesses require additional information about the accuracy of eyewitness memory in order to fully utilise the effortfulness cue. As explained in Chapter 1, warnings about potential memory errors have sometimes been found to reduce the DRM false recognition effect (McCabe & Smith, 2002; Starns et al., 2007) and the misinformation effect (Christiaansen & Ochalek, 1983; Echterhoff et al., 2007; Oeberst & Blank, 2012; Szpitalak

& Polczyk, 2010). For example, Starns et al. (2007) warned participants about the associative nature of DRM lists after participants had studied the lists and informed them that people often falsely recognise critical theme words as having been studied. This warning reduced the DRM false recognition effect, increased the participants' ability to discriminate between studied words and critical theme words, and induced a more conservative response bias (i.e., more withholding). While much of the reduction in the DRM false recognition effect was accounted for by improved discrimination, response bias also played a significant, albeit smaller, role. Thus, it may be important to alter response bias so that people can maximise the quantity and accuracy of their eyewitness memory reports. For example, consider a witness whose average confidence in incorrect memories drops from 70% to 60% in response to mnemonic cue instructions. If this witness sets their response criterion at 60%, the improvement in monitoring will have little impact on the false alarm rate because many incorrect memories will still be above the criterion. It may be that making witnesses more aware of the limitations of their memory will prompt them to consider the accuracy of their memories more carefully and allow them to make better use of the effortfulness cue when they receive instructions about it. This, in turn, should lead them to withhold more inaccurate information because effortfulness appears to be a reliable predictor of accuracy that witnesses do not utilise fully without instruction. Thus, in addition to informing witnesses about effortfulness, they may also need to be explicitly informed that they are likely to recall memories that are incorrect.

Evidence suggests that providing information about the limitations of memory accuracy may be effective because people often have misconceptions about memory, particularly in relation to accuracy. In a study comparing monitoring ability in an eyewitness recall task and a general knowledge test, Perfect (2004, Experiment 2) found that participants believed they would perform better on the eyewitness memory task than

the general knowledge test prior to undertaking the tasks. However, performance was not significantly better in the eyewitness task, and ratings of predictive success were unrelated to performance on the eyewitness task. This finding suggests that witnesses may overestimate their ability to correctly remember details about a crime. The results of Experiments 2 and 3 seem to support this argument given that participants did not withhold information very often, indicating that they may have had an unwarranted level of trust in the information that came to mind. Consistent with this idea, Simons and Chabris (2011) found that 63% of people from a representative sample of the US population believed that human memory works like a video camera and accurately records events for later inspection. Thus, it seems that people may believe that memory is more accurate than it actually is.

In addition, research into judgments of learning has revealed that people's beliefs can affect their metacognitive judgments (Mueller, Tauber, & Dunlosky, 2012). Judgments of learning (JOLs) are predictions about the likelihood of remembering studied information (Dunlosky & Metcalfe, 2009), and people generally give higher judgments (indicating a greater likelihood or remembering) to word pairs that are related as opposed to unrelated, which is termed the relatedness effect (Mueller et al., 2012). Mueller et al. (2012, Experiment 2) examined whether people based their JOLs on beliefs about the relatedness of word pairs. Participants either made their JOLs immediately after studying each word pair, or prior to study. Before making pre-study JOLs, participants were told whether the forthcoming word pair would be related or unrelated. The results showed that pre-study JOLs did produce a relatedness effect, albeit to a smaller degree than immediate JOLs. Thus, people's beliefs about relatedness affected their judgments about the likelihood of remembering word pairs that they had not yet studied. This finding illustrates

that metacognitive judgments can be based, at least partially, on beliefs about memory, meaning that modifying beliefs may improve metacognitive performance.

The low withholding rates observed in Experiments 2 and 3 could also have been due to the type of questions that were used. Accuracy is generally better when questions are open-ended rather than specific and closed (Lipton, 1977), and accuracy rates of more than 90% are often achieved under free-recall conditions (Fisher, 1995). Powell, Fisher, and Wright (2005) have argued that open-ended questioning may allow witnesses to exercise more control over what they volunteer. Thus, the use of closed questions in Experiments 2 and 3 may have inhibited the participants' natural control mechanism, potentially causing them to volunteer incorrect information that they would have withheld under open-ended questioning conditions. For example, participants may set a very liberal response criterion when closed questions are used. This would lead to a greater number of incorrect details being volunteered because a lower level of confidence would be required to justify volunteering an answer, and answers associated with lower confidence are less likely to be correct (Koriat & Goldsmith, 1996). Witnesses may set a more liberal response criterion when answering closed questions than open-ended questions because they feel more obliged to volunteer information. Indeed, evidence from Ackerman and Goldsmith (2008) suggests that people do avoid leaving a large number of questions unanswered during general knowledge tests in which the questions are closed. As closed questions are specific and direct, instances of withholding are obvious because the witness must explicitly state that they cannot remember or do not know the answer. Conversely, as open-ended questions are general and allow for extended responses about a variety of details, instances of withholding are subtle (perhaps unnoticeable) as it is likely that at least some information could be provided in response to such questions. Thus, closed questions may cause witnesses to withhold less information than open-ended questions because they

feel uncomfortable withholding when closed questions are used. This discomfort may stem from the maxim of quantity proposed by Grice (1975) whereby the social norms of conversation dictate that people should provide as much information as possible.

It is also possible that the closed questions used Experiments 2 and 3 inhibited the participants' ability to distinguish between correct and incorrect memories (i.e., monitoring), or that the closed questions simply prompted the retrieval of more inaccurate information than open-ended questions would have. Monitoring ability may have been impaired because closed questions allowed incorrect details to come to mind more quickly and easily than they would have if open-ended questions had been used. Specifically, the presence of the retrieval cues that were not available during open-ended questioning may have provided easier access to the memory, creating an illusion of fluency. This, in turn, may have increased the participants' confidence in these incorrect details to a level above their response criterion, causing them to be volunteered. The closed questions may have also increased the amount of incorrect information being retrieved because some of them asked about things that participants did not remember well. Thus, some of the questions would have been quite difficult. However, other questions may have been quite easy and the task itself may have been considered reasonably simple because the questions were specific and the participants knew exactly what information to search for in their memories. As people's metacognitive judgements are not as sensitive to changes in task difficulty as they could be (Suantak, Bolger, & Ferrell, 1996), the participants may have volunteered incorrect information because they did not adequately account for the difficulty of some of the questions. Specifically, the participants may have set a lenient response criterion due to their perception of the task and the presence of some easy questions, but may not have raised this criterion for difficult questions where they may have been overconfident.

Overall, the findings of Experiments 2 and 3 suggested that providing witnesses with instructions about mnemonic cues does not improve their monitoring ability or impact on the quantity or accuracy of their eyewitness memory reports. It seemed likely that the reasoning instructions were unsuccessful because this mnemonic cue was not found to be a reliable predictor of response accuracy. However, this explanation did not hold for the effortfulness instructions as the effortfulness cue was found to be a reliable predictor of response accuracy. Instead there was an indication that witnesses may need to be informed of the likely inaccuracy of their memory so that they can make more effective use of the effortfulness cue. The aim of Experiment 4 was to explore this possibility. It was also possible that the closed questions used in Experiments 2 and 3 interfered with the natural retrieval, monitoring and/or control mechanisms of participants. Thus, Experiment 5 assessed the impact of question type (open-ended or closed) on these memory regulation processes.

CHAPTER 5 – IMPROVING MONITORING: MNEMONIC CUE KNOWLEDGE, MEMORY INACCURACY WARNING AND RETRIEVAL INSTRUCTIONS

Experiment 4

Experiment 4 explored whether eyewitness monitoring could be improved using a training technique combining instructions about effortfulness with a warning regarding the retrieval of incorrect memories. The memory inaccuracy warning was included because the low withholding rates observed in Experiments 2 and 3 could indicate that witnesses have an unwarranted level of trust in the accuracy of their memory. Indeed, research has suggested that people often have misconceptions about memory accuracy (Perfect, 2004; Simons & Chabris, 2011), and that metacognitive judgments can be based upon beliefs about memory (Mueller et al., 2012). Thus, it seemed possible that altering beliefs about the accuracy of eyewitness memory could change the way people monitor the accuracy of their memory. Warning witnesses about potential inaccuracies in their memory may prompt them to adopt a more conservative response criterion, as has been observed in research on the DRM paradigm (Starns et al., 2007) and misinformation effect (Echterhoff et al., 2007; Szpitalak & Polczyk, 2010). Such a warning could increase accuracy on its own if the more conservative response criterion results in increased withholding of incorrect details. However, correct details may also be withheld which would reduce quantity. If the memory inaccuracy warning is combined with information about effortfulness, this outcome may be avoided as the warning could increase usage of effortfulness. As effortfulness has been found to be a reliable predictor of response accuracy (Experiments 2 and 3), increasing usage of this mnemonic cue should help witnesses withhold more incorrect memories without also withholding correct memories. Thus, the aim of combining a memory inaccuracy warning with effortfulness instructions

was to increase withholding and improve monitoring simultaneously, allowing people to maximise the quantity and accuracy of their eyewitness memory reports.

I also chose to include an experimental condition in which the effortfulness instructions were incorporated into a broader set of instructions regarding how to go about retrieving information from memory. The retrieval instructions were based on a study conducted by Scoboria et al. (2014) which tested the impact of a brief training procedure on responses to answerable and unanswerable questions (i.e., questions about information that was present and not present in the stimulus video, respectively). Scoboria et al. (2014) instructed participants to: (i) review the question, (ii) retrieve all possible responses, (iii) consider the source of the possible answers, (iv) reflect on the likely accuracy of each possible response, (v) and select the best response (i.e., one of the possible answers or *don't know*). Their results showed that while the trained condition achieved higher accuracy for unanswerable questions than the untrained condition, accuracy for answerable questions did not differ significantly between them.,. Although these findings suggest that a brief training procedure may not be useful for improving the way people respond to answerable questions, it could be that such instructions are more useful in situations where ambiguous stimuli are present. Specifically, encouraging thoughtful consideration of retrieved memories may increase awareness of the presence of schema-based intrusions. Therefore, as the stimulus video used throughout this thesis contained a variety of ambiguous stimuli, I chose to incorporate the effortfulness instructions into a training procedure similar to the one used by Scoboria et al. (2014).Experiment 4 consisted of the two experimental conditions (i.e., warning + effortfulness instructions and warning + effortfulness and retrieval instructions) and a control condition that did not receive a memory inaccuracy warning or any effortfulness or retrieval instructions.

I predicted that the experimental conditions would have better monitoring (i.e., type-2 discriminability) than the control condition. I also predicted that the experimental conditions would have a more conservative response bias than the control condition. However, as explained in Chapter 4, the impact of any improvements in monitoring on quantity and accuracy could manifest in different ways. For accuracy to be improved there needs to be (i) an increase in the amount of correct information being volunteered, (ii) a decrease in the amount of incorrect information being volunteered, or (iii) a combination of both. Each of these options constitutes an improvement in monitoring. Thus, if the experimental conditions exhibited better monitoring than the control condition, accuracy was also expected to be higher for the experimental conditions than for the control condition. However, to avoid reducing quantity, the proportion of retrieved information that is both correct and volunteered must be maximised. This means that I only expected quantity to be higher for the experimental conditions if they volunteered a higher amount of correct information than the control condition.

Method

Participants. Ninety participants[37] (62 females, 28 males) took part in the study for course credit or payment ($15). All had normal or corrected-to-normal vision and spoke English as their first language. Participants ranged in age from 17 to 49 (M = 23.13, SD = 7.23).

[37] Before data collection began I decided to recruit 90 participants, 30 per condition. I ended data collection in the week that I reached 80 participants.

Materials.

Stimulus video and recall task. Participants viewed the same video that was used in all previous studies and completed the same recall task that was used in Experiments 2 and 3.

Memory inaccuracy warning. Participants were warned that witnesses retrieve more incorrect information than they think they do, rather than being provided with a more general warning about the retrieval of incorrect information. Although, participants usually chose to volunteer information in Experiments 2 and 3, instances of withholding did occur. This indicated that participants were aware that their memory was wrong on occasion because I would have expected them to volunteer everything if they believed their memories were always correct. Therefore, simply telling witnesses that they will retrieve some incorrect information may not be effective because they are already aware that this occurs.[38] They may discount such a warning and assume that they are correctly identifying all of the instances where their memory is wrong. Therefore, it is important to make them aware that they will underestimate the amount of incorrect information they retrieve from memory.

The underestimation was also framed as being inevitable as opposed to being a possibility. Blank & Launay (2014) conducted a meta-analysis on studies using post-misinformation warnings (i.e., warnings given after the misinformation has been presented) to reduce the misinformation effect. Although they distinguish between four different levels of warning specificity (i.e., possibility, presence, logic of opposition, and identification), only two are relevant to the present study because the others apply

[38] It is also possible that they were only withholding on occasions where they were guessing the answer. If this is the case, a general warning about the retrieval of incorrect information may be effective. Thus, the specific warning would be no less effective yet still have the chance of further benefit.

exclusively to situations in which witnesses receive misinformation. Possibility warnings inform participants that some of the information they received after watching the stimulus video might have been incorrect, while presence warnings state that some of the information was definitely incorrect (Blank & Launay, 2014). Although Blank and Launay (2014) found that presence instructions were not significantly more effective at reducing the misinformation effect than possibility instructions, they noted that this could have been due to the limited number of studies available for the meta-analysis, and/or idiosyncrasies between the studies that were included. It is also possible that they failed to find a significant difference due to demand characteristics. Specifically, participants know that the experimenter has put together the research materials and will therefore assume that the experimenter knows whether the information provided contains inaccuracies. As a result, participants may interpret possibility instructions as presence instructions. However, such a bias is much less likely to occur in the context of eyewitness response accuracy more generally (i.e., in the absence of misinformation). A possibility instruction regarding general eyewitness response accuracy would involve telling witnesses that they might retrieve more incorrect information than they think, while a presence instruction would involve telling them that they will always retrieve more incorrect information than they think. As witnesses generally seem to trust their memory, they may be more likely to discount the possibility that their memory is less trustworthy than they believe it to be. Therefore, in order to minimise the chance that the participants were discounting the memory inaccuracy warning, presence instructions were used instead of possibility instructions.

The memory inaccuracy warning also provided participants with some information about why memory inaccuracies could occur. In their meta-analysis, Blank and Launay (2014) considered the impact of enlightenment warnings on the misinformation effect.

Such warnings inform participants that misinformation was present and explain why (i.e., the scientific motivation and logic behind the misinformation manipulation). Blank and Launay's (2014) results showed that when enlightenment warnings were provided, there was no longer a significant misinformation effect. Blank and Launay (2014) argued that warnings which do not provide further background information may induce scepticism or reluctance in participants. Witnesses may also exhibit reluctance to believe that their memory is wrong given how important memory is for adaptive functioning. Therefore, participants were provided with some common reasons for the occurrence of memory errors in an eyewitness context.

Retrieval and effortfulness instructions. As explained in the introduction of this chapter, the instructions given were similar to those used by Scoboria et al. (2014). They instructed their participants to (i) review each question, (ii) retrieve all possible responses, (iii) consider the source of the possible answers, (iv) reflect on the likely accuracy of each possible response, and (v) select the best possible response (i.e., one of the possible answers or *don't know*). Each of these elements was included in the instructions that were given to the participants, though effortfulness instructions were used in place of the source instructions. Participants were told to review each question before trying to retrieve an answer and were informed that more than one answer could come to mind. In addition, they were instructed to make a decision about the likely accuracy of each possible answer that came to mind by considering effortfulness, and to select the answer that they believed was most likely to be correct. As in Experiment 3, participants were told that memories that are effortful to retrieve are likely to be incorrect and that they should respond *don't know* when their answer to a question was effortful to retrieve.

In contrast to Scoboria et al. (2014) participants were not provided with a printed list of the instructions to refer to during the recall task because, as every other aspect of the

study was conducted on computer, they may have forgotten to refer to it. Instead, the instructions were restated briefly on-screen at several points throughout the questions to ensure that the participants were keeping the instructions in mind throughout the experiment.[39]

Confidence scale. As in Study 1 and Experiment 2, all participants rated their confidence in each answer using a 0% to 100% scale with 10% intervals (i.e., 0%, 10%, 20%, etc.).

Effortfulness rating. Ratings of effortfulness were completed in the same way as all previous experiments by participants in the experimental conditions only. As explained in the methodology section of Experiments 2 and 3, these ratings constituted a second part of the mnemonic cue manipulation as they served as a reminder to the participants that they should consider effortfulness throughout the recall task.

Manipulation check questions. Three manipulation check questions were included to assess whether participants remembered and understood the effortfulness instructions and the memory inaccuracy warning. The first two questions asked participants to rate how important it was to consider effortfulness and the likelihood that their memory was incorrect when providing an eyewitness memory report on a five-point rating scale ranging from 0 (not at all important) to 4 (extremely important). The third question asked participants to rate the extent to which witnesses are likely to remember incorrect

[39] It was not possible to use reaction time data to examine the extent to which participants engaged in the specified retrieval activities. Reactions time data are subject to a large amount of individual variance meaning that results will be unreliable. In retrospect, it would have been beneficial to include several manipulation check question to assess whether participants could recall the retrieval instructions they were given. While this would not have assessed whether they were actually engaging in the retrieval activities during the questions, it would have provided some indication that they were aware of them.

information on a five-point rating scale ranging from 0 (not at all likely) to 4 (extremely likely).

Procedure. After giving informed consent, participants were randomly assigned to one of three levels of the only manipulation: training (i.e., warning and effortfulness instructions, $n = 29$), training + retrieval (i.e., warning and effortfulness instructions included within retrieval instructions, $n = 31$), or control ($n = 30$). Participants were then shown the stimulus video and completed the maze filler task that was used in all previous studies. The time allowed for the filler task was determined by pilot testing the number of minutes required to read each set of instructions. A pilot sample of 20 participants who did not take part in the full experiment were randomly assigned to read one of the three sets of instructions ($n = 7$ for each set of experimental instructions, and $n = 6$ for the control instructions). The findings showed that the instructions took a mean of 2.62 ($SD = 0.33$), 2.68, ($SD = 0.47$), and 4.20 ($SD = 1.44$) minutes to read in the control, training and training + retrieval conditions, respectively. Thus, on average it took an extra minute for participants to read the training + retrieval instructions. Therefore, to equalise the delay between the presentation of the video and the recall questions, participants in the control and training conditions were given 11 minutes to complete the maze task, while participants in the training + retrieval condition were given 10 minutes. All participants completed the experiment on a computer.

After the filler task, participants in training and training + retrieval conditions received a memory inaccuracy warning and instructions about the effortfulness cue (the exact wording of all instructions is provided in Appendix E). For participants in the training + retrieval condition, the effortfulness instructions were incorporated into a set of instructions regarding how to go about retrieving information from memory and deciding how to respond. Participants in the control condition did not receive a memory inaccuracy

warning, effortfulness instructions, or any specific instructions regarding how to go about retrieving answers. However, the importance of accuracy was emphasised to participants in the control condition. After receiving their instructions, all participants completed a practice question (also included in Appendix E). Following the practice question, an abbreviated reminder of the instructions was provided.

The recall task was presented in the same two-phase format used in Experiment 3. When presented with filter questions, participants had two response options, *yes* and *no*. When they said *yes*, they were presented with a closed follow-up question. Participants were asked to think of their best possible answer before proceeding. Those in the training and training + retrieval conditions then rated effortfulness, and all participants rated confidence. They were then asked to decide whether to report the answer or respond *don't know*. When they indicated that they wanted to provide the answer, participants were presented with the question again with a text box to type their answer into. An abbreviated version of the instructions was repeated after the questions about the first offender and again after the questions about the second offender (immediately prior to the final questions regarding how the offenders escaped).

In the second phase of the recall task, participants were asked to answer the closed questions to which they responded *don't know* in phase one. Each of these questions was presented on a separate screen with a text box for typing an answer. Participants were asked to think back to the best possible answer they had in mind when they were first presented with the question. Before being presented with these questions, participants were told that saying *don't know* when they are unsure is good, and that the purpose of asking for responses to these questions was to get an idea of what their best guess was. When this second questioning phase was finished, participants completed the manipulation check questions.

For each question in the recall task I measured only: (i) the participant's response, (ii) the score they gave on the effortfulness rating scale (training and training + retrieval conditions only), (iii) their confidence rating, (iv) whether they chose to volunteer an answer or say *don't know* (closed questions only), and (v) the reaction time for each of these responses. Responses to the manipulation check questions, and the reaction times of these responses were also recorded (none of the reaction time data were analysed).

Results

The relationship between effortfulness and response accuracy. As in Experiments 2 and 3, I began by assessing whether effortfulness predicted response accuracy by constructing a logistic mixed effects model. The results showed that effortfulness was a significant predictor of response accuracy, b = -0.33, SE_b = 0.15, [-0.63, -0.03], and adding it to the model significantly improved fit, χ^2 (1) = 4.03, p = .045. Thus, consistent with the findings of Experiments 2 and 3, effortfulness was found to be a reliable predictor of response accuracy.

Manipulation checks. To assess the success of the warning manipulation, I compared responses to the manipulation check questions that asked participants how important it is to consider response accuracy when providing an eyewitness memory report and how likely witnesses are to retrieve incorrect information.[40] The results of a *t*-test showed that participants who received the warning believed that witnesses were significantly more likely to retrieve incorrect information (M = 2.83, SD = 0.91) than participants who did not receive the memory inaccuracy warning (M = 2.23, SD = 0.86), $t(89) = 3.02$, p = .003, Cohen's d = 0.68. However, a second *t*-test revealed that there was no significant difference in how important participants thought it was to consider response

[40] A mixed effect analysis was not required for these data because there was a single observation for each participant for all items.

accuracy when providing an eyewitness memory report between participants who received the warning (M = 3.33, SD = 0.77) and participants who did not receive the warning (M = 3.57, SD = 0.57), $t(89) = 1.46$, $p = .147$, Cohen's d = -0.33. As both means were on the *extremely important* side of the scale (i.e., above 2), it appeared participants were generally aware that it is important to consider the accuracy of their memory when providing an eyewitness memory report, regardless of whether they received a memory inaccuracy warning. However, it is possible that the combination of participants believing they are less likely to be correct and believing that accuracy is important to consider when providing an eyewitness memory report could have resulted in more withholding. However, the assessment of response bias that is reported in the next section suggested that this was not the case. Thus, it appeared that the memory inaccuracy warning was ineffective.

Next I examined whether the effortfulness instruction manipulation was successful by assessing whether effortfulness predicted control decisions and whether the effortfulness instructions changed how important participants thought it was to consider effortfulness when providing an eyewitness memory report. The results of the logistic mixed effects model showed that effortfulness was a significant predictor of control decisions, b = -1.66, SE_b = 0.19, [-2.04, -1.29], and adding it to the model significantly improved fit, χ^2 (1) = 58.94, $p < .001$. However, a t-test showed that there was no significant difference in how important participants thought it was to consider effortfulness when providing an eyewitness memory report between participants who received effortfulness instructions (M = 3.02, SD = 0.89) and participants who did not receive effortfulness instructions (M = 3.27, SD = 0.74), $t(89) = 1.32$, $p = .189$, Cohen's d = -.30.[41]

[41] A mixed effect analysis was not required for these data because there was a single observation for each participant for all items.

These contradictory results meant that I was not confident regarding the success of the effortfulness manipulation. However, due to the inconsistent results I observed for the manipulation checks, and because I also included the retrieval instructions in one experimental condition, I decided to continue with my assessment of monitoring, response bias, accuracy, and quantity.

The effect of training type on monitoring and response bias. To assess the impact of instruction type and monitoring (i.e., type-2 discriminability), a logistic mixed effects model was constructed that compared the training and training + retrieval conditions to the control condition. This analysis also allowed me to assess the impact of the instructions on response bias. The results of the analysis are presented in Figure 4 and the coefficients are presented in Table 25. As explained in Chapter 2, monitoring is represented by the relative difference between the bars for correct and incorrect responses and response bias is represented by the combined height of the correct and incorrect bars.

The results showed that monitoring did not differ significantly between the control condition and the experimental conditions. Figure 4 shows that the relative difference between the bars for correct and incorrect responses is similar across conditions; though the difference is slightly larger for the training condition and slightly smaller for the training + retrieval condition. Consistent with this, the coefficient for the training condition was positive, indicating that this condition had more effective monitoring than the control condition, though not significantly so. In addition, the negative coefficient for the training + retrieval condition indicated that this condition had less effective monitoring than the control condition, though not significantly so. The lack of a significant difference in monitoring was confirmed by the model fit comparison which showed that adding the interaction between instruction type and response accuracy to the model did not significantly improve fit, χ^2 (2) = 3.46, p = .178.

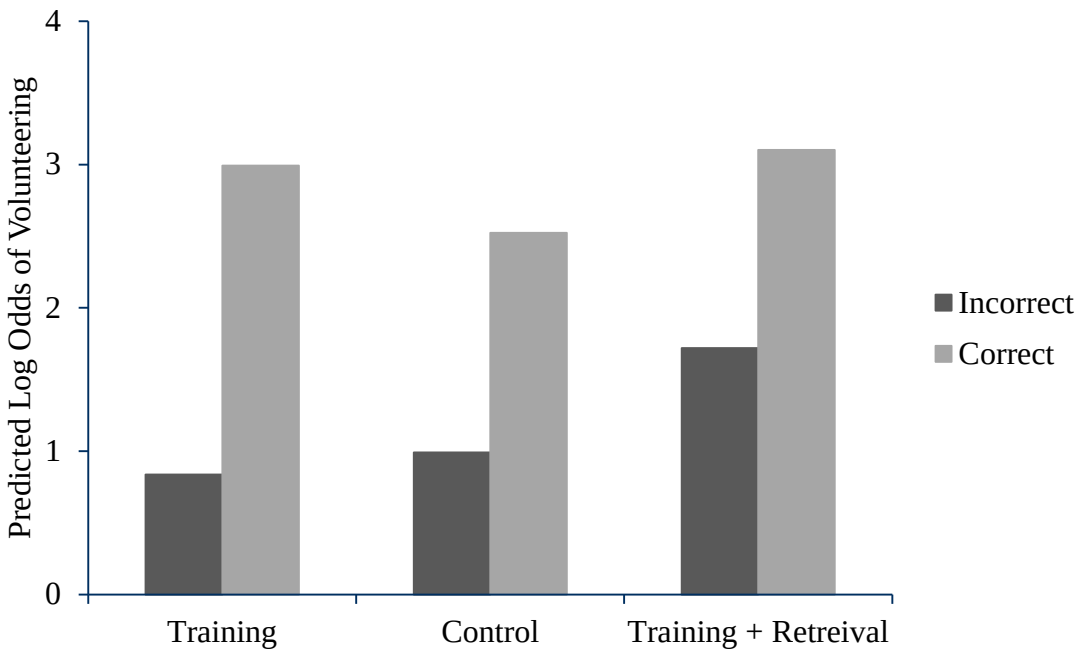

Figure 4. Predicted log odds of volunteering correct and incorrect answers as a function of the type of instructions provided in Experiment 4.

The results also showed that response bias did not differ significantly between the control condition and the training condition or the training + retrieval condition. Figure 4 shows that the combined height of the correct and incorrect bars is similar across conditions; though it is slightly higher for the training and training + retrieval conditions than for the control condition. Indeed, the coefficients for response bias were positive. This indicated that the training and training + retrieval conditions were more likely to volunteer than the control condition, though not significantly so. The lack of a significant difference in response bias was confirmed by the model fit comparison which showed that adding condition to the model did not significantly improve fit, χ^2 (2) = 2.64, p = .268.

Table 25

Fixed Effects Coefficients for Response Bias and Monitoring for the Training and Training + Retrieval Conditions in Comparison to the Control Condition in Experiment 4

	Response bias		
Condition	*b*	SE_b	95% CI_b
Training	0.01	0.36	-0.70, 0.71
Training + retrieval	0.54	0.37	-0.18, 1.27
	Monitoring		
Training	0.62	0.41	-0.18, 1.42
Training + retrieval	-0.15	0.42	-0.97, 0.67

The effect of training type on accuracy and quantity. As there was no evidence that monitoring was significantly improved (or impaired) as a result of the training, I did not expect any significant difference in accuracy or quantity between the conditions, despite my initial predictions. Consistent with this, the results of a logistic mixed effects model showed that accuracy did not differ significantly between the control condition and the experimental conditions (see Table 26 for coefficients and Table 27 for estimated proportions). The lack of a significant difference in accuracy was confirmed by the model fit comparison which showed that adding condition to the model did not significantly improve fit, χ^2 (2) = 1.54, p = .463. Similarly, the results for another logistic mixed effects model showed that quantity did not differ significantly between the control condition and the experimental conditions. The lack of a significant difference in quantity was confirmed by the model fit comparison which showed that adding condition to the model did not significantly improve fit, χ^2 (2) = 0.26, p = .877.

Table 26

Fixed Effect Coefficients for Accuracy and Quantity for the Training and Training + Retrieval Conditions in Comparison to the Control Condition in Experiment 4

	Accuracy		
Training type	*b*	SE_b	95% CI^b
Training	-0.30	0.30	-0.89, 0.30
Training + retrieval	-0.37	0.31	-0.98, 0.24
	Quantity		
Training	-0.15	0.29	-0.71, 0.41
Training + retrieval	-0.08	0.28	-0.64, 0.48

Table 27

Estimated Proportions for Accuracy and Quantity for Each Condition in Experiment 4

Condition	Accuracy	Quantity
Control	.79	.55
Training	.74	.51
Training + retrieval	.73	.53

Discussion

The results of Experiment 4 showed that providing people with a memory inaccuracy warning in addition to effortfulness instructions does not induce a more conservative response bias, improve their monitoring ability, or help them maximise the quantity and accuracy of their eyewitness memory reports Similarly, providing witnesses with information about how to engage in retrieval as well as warning them about memory

inaccuracy and instructing them to use effortfulness does not significantly affect monitoring, response bias, quantity, or accuracy. These findings may be explained by the fact that the warning and effortfulness manipulations were not entirely successful. The discrepancies amongst the manipulation checks are discussed in detail below, followed by a potential explanation for the ineffectiveness of the retrieval instructions.

The discrepancy between the manipulation checks for the memory inaccuracy warning may have been a result of the instructions given to the control condition. Although the warning increased the participants' awareness of the fact that witnesses remember more incorrect information than they think, the participants did not rate the possibility of memory inaccuracy as more important to consider when providing an eyewitness memory report than the control condition did. However, the group means for the question that asked how important it was to consider memory inaccuracy when providing an eyewitness memory report were close to the maximum possible score on the scale (i.e., 4). Thus, participants seemed to believe that considering memory inaccuracy was very important, regardless of the whether they received a warning. This may have occurred as a result of the instructions given to the control group as they emphasised the importance of providing an accurate memory report and strongly discouraged participants from volunteering any information they thought might be incorrect. Thus, the instructions given to the control condition could have increased the participants' perception of the importance of considering memory inaccuracy in a similar way to the warning, masking any impact of the warning. However, if this were the case, I would have expected a higher withholding rate in Experiment 4 in comparison to Experiments 2 and 3 because both the warning and control instructions should have induced a more conservative response bias. This was not the case as the mean withholding rate for Experiment 4 (M = 0.16, SD = 0.13) was very similar to the withholding rates observed in Experiments 2 (M = 0.16, SD = 0.14)

and 3 (M = 0.19, SD = 0.13). This comparison suggests that the warning manipulation was indeed unsuccessful, which explains why it did not have a significant impact on monitoring, quantity, or accuracy. It is possible that people may require more intensive education regarding the accuracy of memory in order to alter their beliefs about it. Indeed, Lane and Karam-Zanders (2013) have argued that changing people's knowledge about their memory may be quite difficult and will require more than mere exposure to accurate factual information.

The fact that the memory inaccuracy warning appeared to have been unsuccessful may also explain why the effortfulness instructions did not affect monitoring, quantity, or accuracy, though it is also possible that the effortfulness manipulation itself was unsuccessful. As explained in Chapter 4 participants may not make proper use of the effortfulness cue if they believe their memory is usually correct because they see little reason to use a mnemonic cue that helps identify incorrect memories. Based on this, I argued that increasing awareness of the limitations of memory may prompt increased usage of effortfulness because it should make witnesses more willing to consider the possibility that their memory could be incorrect. As the memory inaccuracy warning I used in Experiment 4 did not appear to increase awareness of memory limitations, it is unsurprising that the effortfulness instructions were ineffective. However, there was some indication that the manipulation of effortfulness was unsuccessful which could also explain the findings. Specifically, participants who received effortfulness instructions did not report believing it was more important to consider effortfulness when providing an eyewitness memory report than participants who did not receive effortfulness instructions. This finding could indicate that people consider effortfulness when providing an

eyewitness memory report without being told to do so.[42] Consistent with this, the means for both groups were on the side of the scale that indicated effortfulness was thought to be quite important (i.e., above 2). Thus, witnesses may spontaneously consider effortfulness during monitoring.

There were several differences between the methodology used in Experiment 4 and the methodology used by Scoboria et al. (2014) which may explain why the retrieval instructions were ineffective. Scoboria et al. (2014) demonstrated their retrieval instructions using an example which may have improved understanding of the instructions. Consequently, this may have resulted in more successful application of the retrieval technique, leading to improved performance (i.e., more correct rejections of unanswerable questions). Thus, as the participants in Experiment 4 were not provided with a specific example, their understanding and usage of the retrieval technique may have been lacking, which could have impacted on the success of the technique. Understanding and usage of the retrieval technique in Experiment 4 may have also been reduced because participants were not asked to verbally articulate their retrieval process during the practice question. Scoboria et al. (2014) had participants verbally articulate their retrieval process to an interviewer who was able to monitor the participants' understanding of the instructions and answer any questions. This process may have been crucial for understanding and applying the retrieval technique. Thus, that lack of a verbal articulation procedure in Experiment 5 may also explain why the retrieval instructions were ineffective.

[42] Of course, it is also possible that the manipulation check responses reflected demand effects. Specifically, participants may have been saying what they thought was expected of them. However, the lack of objective metacognitive and memory differences between groups suggest that demand effects did not simply mask an otherwise effective manipulation.

It is also possible that the discrepancy between my findings and those of Scoboria et al. (2014) can be explained by the fact that they used explicitly unanswerable questions (i.e., questions about information that was not present in the stimulus video). Specifically, 10 of the 20 interview questions asked by Scoboria et al. (2014) were unanswerable, meaning that *don't know* was the correct response for half of the questions. The fact that there were so many unanswerable questions may have instilled a belief that *don't know* responses were actually helpful because participants were in a state where they were aware that they did not know. In contrast, the majority of questions used in Experiment 4 could be answered if the participants had been able to encode everything.[43] This may have instilled a belief in participants that they should know the answer to most questions, meaning that they may have perceived the *don't know* response as unhelpful. Ultimately, this may have led to a reluctance to withhold, a reluctance that might be reduced when there is a high proportion of objectively unanswerable questions, as in Scoboria et al.'s (2014) study.

In conclusion, Experiment 4 found no evidence that eyewitness monitoring can be improved by providing a memory inaccuracy warning, effortfulness instructions, or retrieval instructions. The withholding rate was very low, and similar to that rate observed in Experiments 2 and 3, with the memory inaccuracy warning being unable to induce a more conservative response bias. As the findings suggested that the warning and effortfulness manipulations were unsuccessful, more effective means of manipulating knowledge about the limitations of memory and knowledge about effortfulness may need to be explored. Further exploration of the effectiveness of retrieval instructions is also

[43] Some closed questions were unanswerable if participants answered a filter questions incorrectly. For example, if a participant said the robber by the counter was wearing a jumper even though they were not, it would not be possible to answer questions about the colour and type of jumper worn.

required as there may have been issues with the comprehension and application of the retrieval technique within Experiment 4.

CHAPTER 6 – CONTROL, MONITORING AND RETRIEVAL OF INCORRECT INFORMATION DURING OPEN-ENDED AND CLOSED QUESTIONING

Experiment 5

Experiment 5 examined whether the low withholding rates observed in Experiments 2-4 were a consequence of the type of questions that were used. In Chapter 4, I explained that the closed questions I used may have inhibited the participants' natural control mechanism, which could have caused them to volunteer incorrect information that they would have withheld if open-ended questions had been used. This explanation was based on Powell et al.'s (2005) assertion that open-ended questions may elicit more accurate memory reports than closed questions because they allow witnesses to exercise more effective control over what they volunteer. However, as explained in Chapter 4, it is possible that the closed questions inhibited the participants' monitoring ability because they allowed incorrect details to come to mind more quickly and easily than they would have if open-ended questions had been used, which may have interfered with the participants' ability to identify the incorrect details. In addition, the closed questions could have simply prompted the retrieval of a greater amount of incorrect information than open-ended questions would have because some of the questions were probably quite difficult to answer, while others may have been quite easy and the task itself may have been considered reasonably simple. As metacognitive judgements are not as sensitive to changes in task difficulty as they could be (Suantak et al., 1996), the participants may have volunteered incorrect information because they did not adequately account for the difficulty of some of the questions. To discover whether the closed questions used in Experiments 2-4 could account for the low withholding rates that were observed, Experiment 5 aimed to (i) compare the control strategies employed during open-ended and closed questioning, (ii) evaluate monitoring ability during open-ended and closed

questioning, and (iii) assess the amount of incorrect information retrieved in response to open-ended and closed questions.

Exploring potential differences in retrieval, monitoring, and control between open-ended and closed questions may assist in developing techniques that allow witnesses to maximise the quantity and accuracy of their eyewitness memory reports when closed questions are used. As explained in Chapter 1, open-ended questions tend to yield more accurate eyewitness memory reports than closed questions (Fisher, 1995; Lipton, 1977). Thus, understanding why this open question accuracy advantage occurs may uncover ways of increasing accuracy in response to closed questions. For example, if monitoring and control strategies are more effective during open-ended than closed questioning, reasons for the difference can be explored. Should the regulation strategies used during open-ended questioning be successfully applied to closed questioning, the accuracy of eyewitness memory reports obtained via closed questions should be increased (though it is important that this increase in accuracy is not accompanied by too great a reduction in quantity). Thus, determining why open-ended questions result in more accurate eyewitness memory reports may help develop ways to improve the way witnesses respond to closed questions.

In order to assess differences in the quality of retrieved information, monitoring, and control between open-ended and closed questions, I required a method that measured withheld responses as well as volunteered responses for open-ended questions. In typical open-ended questioning procedures, participants are asked one or more broad questions (e.g., 'Can you describe the sequence of events?' and 'Can you describe how the offender/s escaped?'), and can give as extensive a response as they choose. The answer they provide may represent everything they were able to retrieve from memory, or it could represent only the subset of retrieved information that they felt comfortable volunteering. Even if

participants are asked to write down everything that comes to mind, it is unknown whether they would have withheld some of the information if given the option. As an explicit distinction between volunteered and withheld information is required to assess monitoring and control, typical open-ended questioning procedures are inadequate. Therefore, to obtain responses for open-ended questions, I used an externalised free-recall technique (Carneiro & Fernandez, 2013; Hollins, Lange, Berry, & Dennis, under review; Hollins, Lange, Dennis, & Longmore, 2015; Kahana, Dolan, Sauder, & Wingfield, 2005; Unsworth, Brewer, & Spillers, 2010) which asked participants to be more liberal in their output and distinguish between details they wanted to volunteer and withhold.

Traditional externalised free-recall tasks ask participants to volunteer everything that comes to mind during retrieval, and to indicate which of the retrieved details are believed to be correct (Carneiro & Fernandez, 2013; Kahana et al., 2005; Unsworth et al., 2010). However, I chose to ask participants to distinguish between responses they wanted to volunteer and withhold because the correct/incorrect method would have required me to assume that responses marked as correct would be volunteered and responses marked as incorrect would be withheld. This would not have been ideal because correct/incorrect distinctions may not always match volunteer/withhold distinctions. For example, a participant may be 70% confident in a particular detail and decide it is correct when forced to a make a 2-alternative accuracy decision (i.e., if they decide to say *correct* for everything above 60%). However, this participant may choose to withhold this information if given the opportunity because confidence is below their report criterion (i.e., if their criterion is set at 80%). Due to this potential problem, it was important to have an explicit measure of control, which was also used in the assessment of monitoring. Thus, using the externalised free-recall task with volunteer/withhold decisions allowed

uninhibited recall while also providing the necessary information to assess monitoring and control for open-ended questions.

Although there are various formats for getting people to indicate which responses they want to volunteer and withhold, the most appropriate for the purposes of Experiment 5 was the two-column method rather than the asterisk or button press method. In the asterisk method, participants mark the responses they want to withhold with an asterisk (Carneiro & Fernandez, 2013). They do this during the recall process rather than retrospectively to allow observation of online monitoring processes (Carneiro & Fernandez, 2013). However, participants cannot be stopped from adding asterisks retrospectively, and there is no way to determine whether this has occurred. One way to prevent this retrospective monitoring is to have participants enter responses via a computer and press a particular key (e.g., the space bar) when they want to withhold a response (Unsworth et al., 2010). However, it is possible that participants could forget about the withhold option and/or they may accidentally press the button for a response they want to volunteer. Another alternative is to provide participants with two columns in which to write their responses (Hollins et al., under review, 2015); a volunteer column and a withhold column. Although it is still possible for participants to retrospectively move an answer from one column to the other in this situation (i.e., by using an arrow or by crossing out the answer and re-writing it in the other column), the instances where this occurs can be recorded. The two-column method also serves as a subtle reminder of the externalised free-recall instructions because both options are always present during recall. Therefore, participants in Experiment 5 were provided with volunteer and withhold columns in the externalised free-recall task.

Two conditions which differed in the extent to which accuracy was emphasised were included to assess adherence to the externalised free-recall instructions. Participants

in the weak accuracy emphasis condition were told that they should try to provide correct information. They were told to write responses in the *volunteer* column when there was a high chance that they were correct. However, participants in the strong accuracy emphasis condition were told it was important to only provide correct information. They were instructed to only write responses in the *volunteer* column when they were absolutely sure they were correct, and were told that if they had any doubt about their memory, they should not use the *volunteer* column. When accuracy is strongly emphasised, participants should write fewer responses in the *volunteer* column and more responses in the *withhold* column than when accuracy is only weakly emphasised. The amount of information retrieved overall should not differ between the conditions if they are using the *withhold* column as instructed.

I predicted that open-ended questions would result in higher accuracy than closed questions. Due to the exploratory nature of Experiment 5, no specific predictions were made about differences between the two question types in terms of the retrieval of correct and incorrect information, monitoring (i.e., type-2 discriminability), or control (i.e., type-2 response bias). However, I expected to observe a difference between the question types for at least one of these dependant measures if a difference in accuracy was apparent.

Method

Participants. Eighty participants[44] (52 females, 28 males) took part in the study for course credit or payment ($10). All had normal or corrected-to-normal vision and spoke English as their first language. Participants ranged in age from 18 to 49 (M = 23.85, SD = 7.92).

[44] Before data collection began we decided to recruit 80 participants, 40 per condition. I ended data collection in the week that I reached 80 participants.

Materials.

Stimulus video and closed question recall task. Participants viewed the same video that was used in all previous studies and they completed the same set of filter and closed questions that were used in Experiments 2-4.

Open-ended question recall task. The open-ended recall questions were administered via an externalised free-recall task described in the introduction of this chapter. There were six questions: (i) 'What can you recall about the sequence of events?', (ii) 'What can you recall about the appearance and clothing of the robber by the counter?', (iii) 'What can you recall about the appearance and clothing of the robber off to the side?', (iv) 'What can you recall about how the robber by the counter escaped?', (v) 'What can you recall about how the robber off to the side escaped?' and (vi) 'Is there anything else you can recall?'.

A coding guide was constructed to code participants' responses as either correct or incorrect. Each discreet piece of information was coded separately. For example, if a participant said one of the robbers was wearing a black balaclava, they would get one code for describing the colour of the disguise and one for describing the type of disguise (the full coding guide provided in Appendix F). Responses relating to the physical (e.g., clothing, build, height etc.), behavioural (e.g., speech and movements/actions), and sequential (e.g., order of robber entry/departure) aspects of the stimulus video were coded. Responses relating to emotions and intentions were ignored because of their subjective and/or speculative nature. Predictions about what may have happened after the recording ended were also ignored. Coding was completed by me and an independent rater on a random sample of 10 open-ended recall tasks. There was an acceptable level of interrater agreement at the item level, $\kappa = .84$ [.83, .87].

Manipulation checks. In addition to the accuracy emphasis manipulation outlined in the introduction of this chapter (i.e., one condition instructed to volunteer responses that had a high chance of being correct and another instructed to only volunteer responses that were definitely correct), a set of eight general knowledge questions were included. These questions were answered using the two-column method to determine whether participants understood and followed the externalised free-recall instructions.[45] Four of the questions related to very basic facts that most participants were expected to answer correctly (i.e., easy questions), while the other four questions related to obscure facts that most participants were expected to answer incorrectly if forced to respond (i.e., difficult questions). These expectations were confirmed by the data with mean proportions correct of .95 (*SD* = .21) and .03 (*SD* = .18) for the easy and difficult questions, respectively. The difficulty of the questions was varied so that both correct and incorrect answers would be retrieved. If participants were adhering to the externalised-free-recall instruction, correct answers should predominantly be written in the volunteer column and incorrect answers should predominantly be written in the withheld column.[46]

Procedure. After giving informed consent, participants were randomly assigned to one of two levels of the accuracy emphasis manipulation: strong emphasis or weak emphasis (n = 40 in each condition). Participants were then shown the stimulus video and completed the 10-minute maze filler task that was used in all previous studies. After the filler task, participants received their accuracy emphasis instructions, completed the open-

[45] In contrast to the open-ended interview questions, the general knowledge questions were presented in cued-recall format.

[46] Alternatively, volunteering and withholding could have been compared between easy and difficult questions. However, I chose to compare volunteering and withholding between correct and incorrect answers because some people may know facts that are, on average, rarely known and vice versa.

ended questions, and then the closed questions. Thus, question type was manipulated within-subjects and the question types were presented in a fixed order.

Each open-ended question was presented on a single piece of paper and participants provided hand-written responses. On each page, the question appeared at the top followed by the instruction 'Please write down everything that comes to mind'. Under the question and instruction were two columns marked *volunteer* and *withhold*. For participants in the strong accuracy emphasis condition, '(absolutely correct)' was written under the *volunteer* heading. Before completing the questions, participants were given detailed instructions on how to use the two columns, including an example unrelated to the stimulus video.

The closed questions were completed on the computer in the same format as Experiments 3 and 4. When presented with filter questions, participants had two response options, *yes* and *no*. When they said *yes*, they were presented with a closed follow-up question. Participants were asked to think of their best possible answer before proceeding. When presented with a closed question, participants were asked to think of an answer before proceeding. They were then asked whether they would like to provide the answer they had in mind or say *don't know*. When they indicated that they wanted to provide the answer, the question was presented again with a text box for typing the answer. When participants had responded to all of the closed questions, they were asked to provide an answer to the questions that had received a *don't know* response. Participants were asked to think back to the answer they had in mind when they were first presented with the question. Before being presented with these questions, participants were told that although saying *don't know* when they are unsure is good; the purpose of asking for responses to these questions was to get an idea of what their best guess was.

When participants had completed both sets of questions, they were asked to answer the general knowledge questions. They were given the list of questions and a separate

sheet of paper with the *volunteer* and *withhold* columns. As in the initial open-ended question recall task, '(absolutely correct)' was written under the *volunteer* heading for participants in the strong accuracy emphasis condition. Participants were told to complete the task in the same way as they had completed the first set of open-ended questions. They were asked to provide answers to all of the general knowledge questions, even if they had to guess. No other instructions were given.

For each open-ended question and the manipulation check task, I measured only: (i) the participant's responses, and (ii) the column in which they recorded each response. For each closed question, I measured only: (i) the participant's response, (ii) whether they chose to volunteer or withhold the answer, and (iii) the reaction time for each of these responses (none of the reaction time data were analysed).

Results

Manipulation checks. It was not possible to properly assess adherence to the externalised free-recall instructions by comparing control decisions and the amount of information retrieved in the weak and strong accuracy emphasis conditions because the accuracy emphasis manipulation was unsuccessful. Specifically, a logistic mixed effects model revealed that the strong accuracy emphasis condition did not achieve significantly higher accuracy than the weak accuracy emphasis condition, $b = 0.47$, $SE_b = 0.26$, [-0.04, 0.98], and adding accuracy emphasis condition to the model did not significantly improve fit, $\chi^2 (1) = 3.37$, $p = .066$. As the accuracy emphasis manipulation did not have a significant impact on accuracy, it was not expected to have had an impact on control decisions. Consistent with this expectation, a logistic mixed effects model showed that the strong accuracy emphasis condition did not withhold significantly more details than the weak accuracy emphasis condition, $b = -0.58$, $SE_b = 0.30$, [-1.16, 0.00], and adding accuracy emphasis condition to the model did not significantly improve fit, $\chi^2 (1) = 3.79$, p

= .051. Thus, although there was no significant difference in the amount of retrieved information between the strong (*M* = 32.95, *SD* = 9.23) and weak accuracy emphasis conditions (*M* = 31.15, *SD* = 8.92), *t*(77.91) = 0.89, *p* = .378, *d* = 0.20, I cannot determine whether participants were following the externalised-free recall instructions based on the accuracy emphasis manipulation.

Fortunately, it was possible to assess adherence to the externalised free-recall instructions by examining data from the general knowledge questions. Responses to the general knowledge questions were counted as being consistent with the externalised free-recall instructions if they were: (i) correct and written in the *volunteer* column, or (ii) incorrect and written in the *withhold* column. Incorrect answers written in the *volunteer* column, correct answers written in the *withhold* column, and non-responses[47] (i.e., when participants wrote 'don't know', 'unsure' or left a blank space) were coded as failures to follow the externalised free-recall instructions. The participants' responses were consistent with the instructions in 82.34% of cases. This finding suggested that participants comprehended and were compliant with the externalised free-recall instructions.[48]

The effect of question type on accuracy, the quality of retrieved information, monitoring and response bias. I ran the analyses on each of two different subsets of the data to assess potential differences between the question types. As the ultimate aim of my

[47] This type of response occurred in only 11.88% of cases.

[48] The general knowledge task was not designed to assess the participants' ability to discriminate between correct and incorrect answers. Rather, it was designed to examine whether the majority of participants wrote down a response for all answers and used the columns as instructed (i.e., by placing answers they believed to be correct or incorrect in the volunteer or withhold columns, respectively). If the task had included moderately difficult questions, it would be impossible to determine whether a participant who placed an incorrect answer in the volunteer column did so because they were unable to follow the instructions or unable to recognise the answer as incorrect.

research was to discover ways to train witnesses to monitor the accuracy of their memory more effectively in response to closed questions, I first assessed performance for items that were referred to in the closed questions. Including items that could only be provided during open-ended questioning (i.e., because there were no closed questions about these items) may not reveal any useful information about how to improve monitoring during closed-ended questioning. For example, monitoring may be better for items that were only provided during open-ended questioning but similar for items that could be provided in both question formats. In this instance, learning more about the monitoring processes operating during open-ended questioning is unlikely to improve the way people respond to closed questions because they monitor the information covered in the closed questions to a similarly effective degree. Therefore, the first set of analyses included items that could be matched across open-ended and closed questions. Specifically, only open-ended items that could be matched to one of the 34 closed questions were included in the first data set. However, this ignores a large amount of the items provided during the open-ended questioning, partially due to the limited number of closed questions that were included. While the selection of closed questions used in the experiments presented in this thesis was based on what police would be likely to ask during interviews, the data obtained during the open-ended questions made it clear that the list of closed questions was not exhaustive. On average, participants provided an extra 15.34 (*SD* = 6.74) details in response to the open-ended questions. Thus, if additional closed questions had been included, it may have been possible to conduct a more balanced assessment of the monitoring processes that operate during open-ended and closed questioning, a point that will be returned to in the discussion section of this chapter. However, a second set of analyses was conducted that included all of the data for the open-ended questions. If monitoring is found to be superior during open-ended questioning in this analysis, it will be important for future studies to assess

monitoring with a larger set of closed questions. The results for the matched data are presented first, followed by the results for the full data set.

Matched data set. I began by assessing whether accuracy differed depending on the type of questions that were used by constructing a logistic mixed effects model comparing open-ended and closed questions. Contrary to expectations, the coefficients revealed that closed questions did not result in significantly lower accuracy than open-ended questions in the matched data set, b = -0.43, SE_b = 0.34, [-1.09, 0.24], and model fit was not significantly improved with the addition of question type, χ^2 (1) = 1.46, p = .227. Although accuracy did not significantly differ between the two types of questioning, it was possible that the two types of questioning achieved similar levels of accuracy via different means. Thus, I continued with my assessment of the possible underlying mechanisms.

Initially, I examined whether there were differences in the quality of retrieved information by considering the amount of correct and incorrect details that were retrieved during open-ended and closed questioning. A two-way repeated measures ANOVA tested whether the number of retrieved items differed between open-ended and closed questions as a function of response accuracy (Figure 5). There was a significant main effect of question type with participants retrieving significantly more items for closed questions than for open-ended questions, F(1, 79) = 175.55, p < .001, η_G^2 = .20. There was also a significant main effect of accuracy with participants retrieving more correct items than incorrect items, F(1, 79) = 423.98, p < .001, η_G^2 = .74. The interaction was also significant, F(1, 79) = 4.13, p = .045, η_G^2 < .01. Simple effects tests revealed that a greater number of correct items were retrieved for closed questions than for open-ended questions, t(157.99) = 4.50, p < .001, d = 0.72. Similarly, a significantly greater number of incorrect items were retrieved for closed questions than open-ended questions, t(135.50) = 8.91, p < .001,

d = 1.42. Thus, closed questions resulted in significantly more correct and incorrect items being retrieved, but the effect was greater for incorrect items.

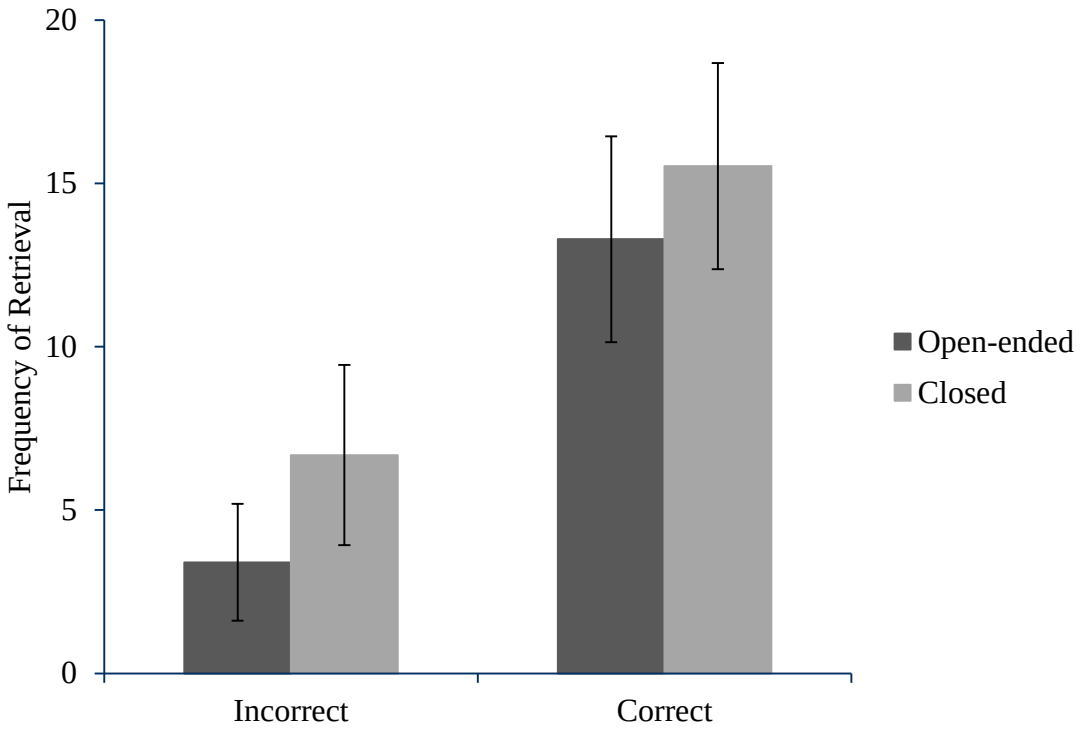

Figure 5. Mean number of correct and incorrect details retrieved in response to open-ended and closed questions in Experiment 5 when items were matched across question type. Error bars represent standard deviations.

Next, I assessed the impact of instruction type on monitoring (i.e., type-2 discriminability) by constructing a logistic mixed effects model comparing closed and open-ended questions. This analysis also allowed me to assess the impact of question type on response bias. The results of the analysis are presented in Figure 6. As explained in Chapter 2, monitoring is represented by the relative difference between the bars for correct and incorrect responses, while response bias is represented in the figure by the combined height of the correct and incorrect bars.

The results revealed that there were significant differences in both monitoring and response bias between open-ended and closed questions. Figure 6 shows that the relative difference between the bars for correct and incorrect responses is smaller for closed questions than for open-ended questions, and the coefficients revealed that participants had significantly less effective monitoring for closed questions than for open-ended questions, b = -0.87, SE_b = 0.38, [-1.60, -0.13]. The significant difference in monitoring was confirmed by the model fit comparison which showed that adding the interaction between question type and response accuracy to the model significantly improved fit, χ^2 (1) = 5.40, p = .020. Figure 6 also shows that the combined height of the correct and incorrect bars is lower for closed questions than open-ended questions, and the coefficients revealed that participants were significantly more likely to withhold for closed questions than for open-

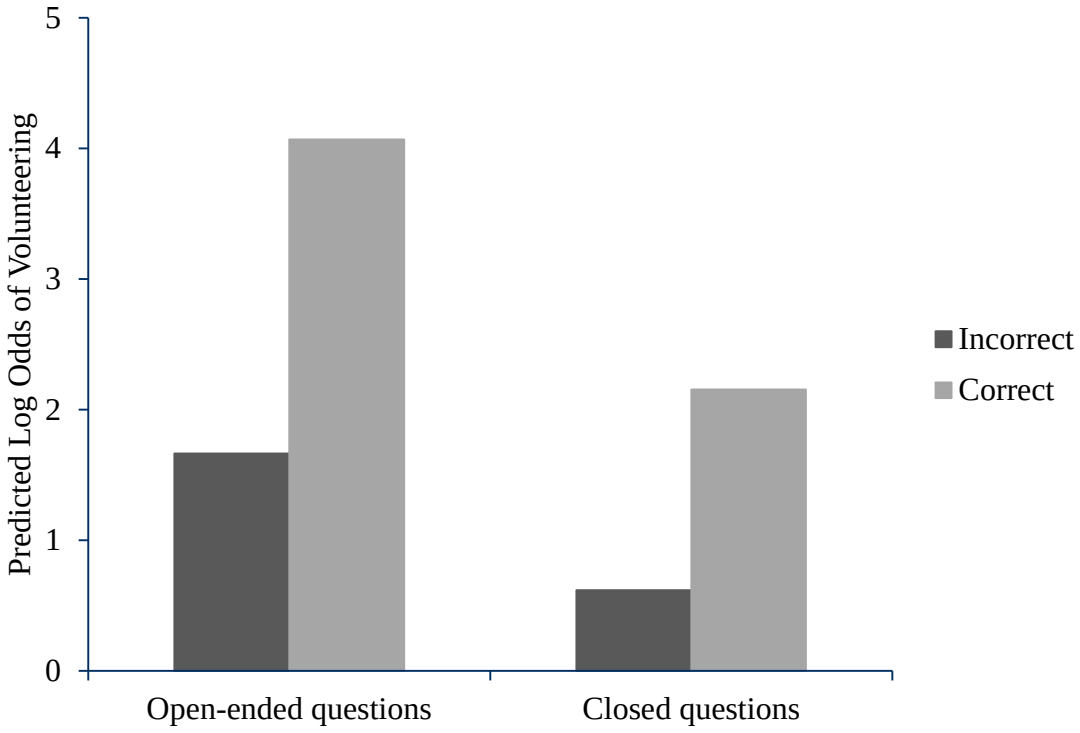

Figure 6. Predicted log odds of volunteering correct and incorrect answers as a function of question type in Experiment 5 when items were matched across questions type.

ended questions, b = -1.19, SE_b = 0.25, [-1.68, -0.71]. The significant difference in response bias was confirmed by the model fit comparison which showed that adding question type to the model significantly improved fit, χ^2 (1) = 17.98, p < .001.

Full data set. Data for the full data set were analysed in the same way as the matched data set. Thus, I began by assessing whether accuracy differed depending on the type of questions that were used by constructing a logistic mixed effects model comparing open-ended and closed questions. As expected, the coefficients revealed that closed questions resulted in significantly lower accuracy than open-ended questions in the full data set, b = -1.08, SE_b = 0.44, [-1.94, -0.21], and model fit was significantly improved with the addition of question type, χ^2 (1) = 5.82, p = .016. Having observed a difference in accuracy, I went on to examine the mechanisms that may underlie this difference.

Initially, I examined whether there were differences in the quality of retrieved information by considering the amount of correct and incorrect details that were retrieved during open-ended and closed questioning. A two-way repeated measured ANOVA tested whether the number of retrieved items differed between open-ended and closed questions as a function of response accuracy (Figure 7). There was a significant main effect of question type with participants retrieving significantly more items for open-ended questions than for closed questions, F(1, 79) = 92.96, p < .001, η_G^2 = .22. There was also a significant main effect of accuracy with participants retrieving more correct items than incorrect items, F(1, 79) = 526.18, p < .001, η_G^2 = .71. The interaction was also significant, F(1, 79) = 224.19, p < .001, η_G^2 = .27. Simple effects tests revealed that while a significantly higher number of correct details were retrieved for open-ended questions compared to closed questions, t(104.94) = 11.49, p < .001, d = 1.83, there was no

significant difference in the number of incorrect details retrieved, $t(151.64) = 1.66$, $p = .099$, $d = 0.26$.

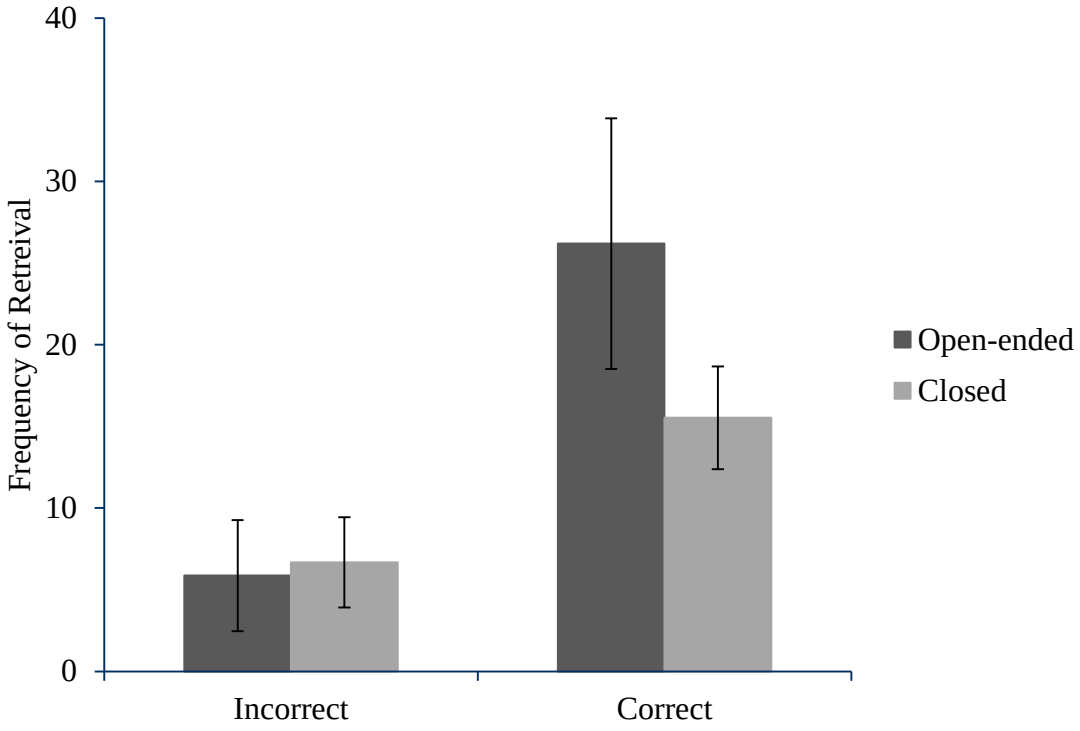

Figure 7. Mean number of correct and incorrect details retrieved in response to open-ended and closed questions in Experiment 5 when all responses to the open-ended questions were included in the analysis. Error bars represent standard deviations.

Next, I assessed the impact of instruction type on monitoring (i.e., type-2 discriminability) by constructing a logistic mixed effects model comparing closed and open-ended questions. This analysis also allowed me to assess the impact of question type on response bias. The results of the analysis are presented in Figure 8. Recall that monitoring is represented by the relative difference between the bars for correct and incorrect responses, while response bias is represented in the figure by the combined height of the correct and incorrect bars.

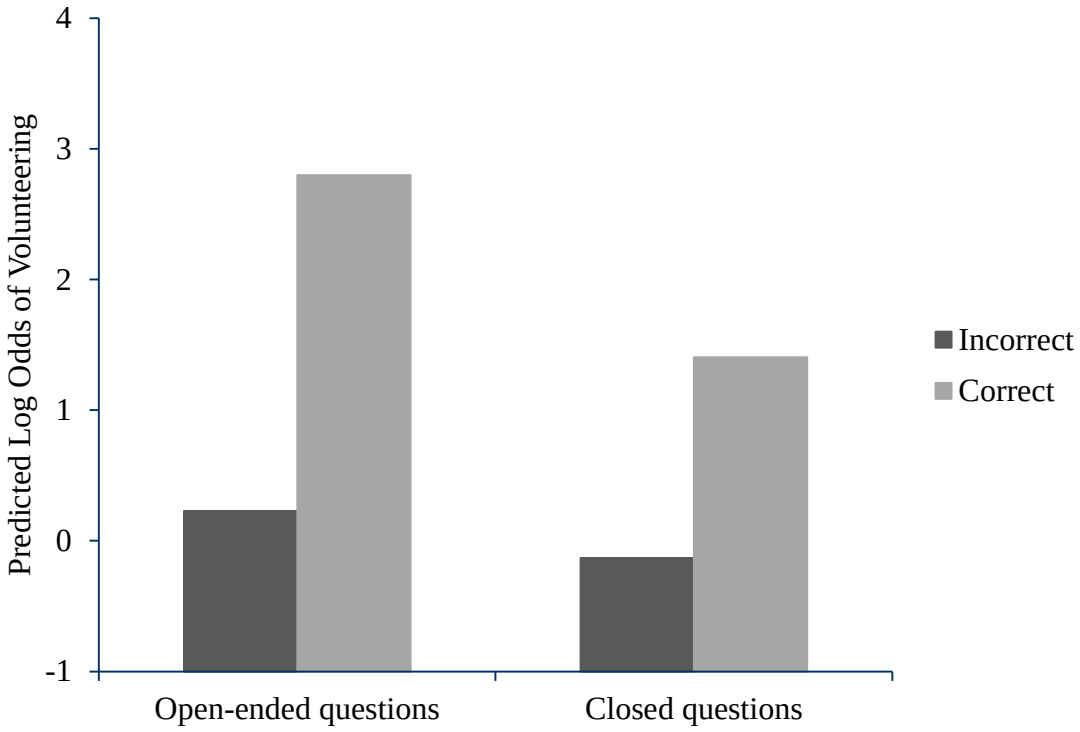

Figure 8. Predicted log odds of volunteering correct and incorrect answers as a function of question type in Experiment 5 when all responses to the open-ended questions were included in the analysis.

The results revealed that there were significant differences in both monitoring and response bias between open-ended and closed questions. Figure 8 shows that the relative difference between the bars for correct and incorrect responses is smaller for closed questions than for open-ended questions, and the coefficients revealed that participants had significantly less effective monitoring for closed questions than for open-ended questions, b = -1.03, SE_b = 0.33, [-1.68, -0.39]. The significant difference in monitoring was confirmed by the model fit comparison which showed that adding the interaction between question type and response accuracy to the model significantly improved fit, χ^2 (1) = 9.55, p = .002. Figure 8 also shows that the combined height of the correct and incorrect bars is lower for closed questions than open-ended questions, and the coefficients revealed that

participants were significantly more likely to withhold for closed questions than for open-ended questions, b = -1.13, SE_b = 0.20, [-1.53, -0.74]. The significant difference in response bias was confirmed by the model fit comparison which showed that adding question type to the model significantly improved fit, χ^2 (1) = 18.83, p < .001.

Discussion

The aims of Experiment 5 were to (i) assess whether the low withholding rates observed in Experiments 2-4 could be explained by the type of questions that were used, (ii) examine the mechanisms that underlie the open-question accuracy advantage observed in the literature, and (iii) explore whether the memory regulation processes that operate during open-ended questioning offer a potential avenue for improving the way witnesses respond to closed questions. The findings suggested that the closed questions used in Experiments 2-4 did not inhibit the participants' natural control mechanism. In addition, the results showed that the open question accuracy advantage may occur due to differences in the retrieval, monitoring, and control mechanisms that operate during open-ended and closed questions. However, based on the results obtained, it is unclear whether the memory regulation processes that operate during open-ended questioning could offer a means of helping people maximise the quantity and accuracy of their eyewitness reports when closed questions are used.

In contrast to the explanation I proposed in Chapter 4 regarding the low withholding rates observed in Experiments 2 and 3 (and subsequently Experiment 4), the results of Experiment 5 suggested that the closed questions used throughout this thesis did not inhibit the natural control mechanism of participants by inducing an extremely liberal response bias. When only considering items that were included in the closed questions used in Experiments 2-5, withholding was higher in response to closed than open-ended questions. This indicates that for the items considered across the experiments presented in

this thesis closed questions may have in fact induced a more conservative response bias than open-ended questions would have. However, it is important to note that this does not necessarily mean that closed questions induce a more conservative response bias in general. Although the results for the full data set also showed that withholding was higher in response to closed than open-ended questions, open-ended questions prompted the retrieval a large amount of additional information. Thus, the more liberal response bias observed for open-ended questions in the full data set may be more reflective of the additional information being retrieved, rather than a stronger tendency to volunteer information.

It is possible that participants in Experiment 5 were able to retrieve and volunteer additional information in response to open-ended questions due to the associative nature of memory. When a witness is asked to provide a description of an offender (as opposed to being asked a variety of closed questions about the offender), they may engage in a general retrieval strategy in which they rely on the information that comes to mind freely (i.e., without a prompt or deliberate search). Such a retrieval strategy may give rise to a series of very strong memories, increasing the witness's confidence and reducing their inclination to withhold information. Indeed, Craik and Tulving (1975) found that deeply encoded words were more likely to be retrieved during a free-recall test than words that are encoded more shallowly. Thus, it is plausible that the open-ended questions resulted in increased retrieval of correct memories because they made participants more likely to retrieve information that was well encoded. This idea aligns well with Anderson's (1976, 1983a, 1983b) associative network model of memory which proposes that information is represented in memory as a network of nodes linked via association. In this model, the quality of encoding determines the strength of associative links. Thus, when a witness retrieves a memory that was well encoded, it will prompt the retrieval of other strongly

encoded memories. As open-ended questions provide very few cues regarding the information that is sought, witnesses can take advantage of the associative networks that exist within memory to retrieve memories that were encoded most strongly. Conversely, when closed questions are used, the information being sought is specified in the question and there is a greater chance of weakly encoded memories becoming activated. In addition, even if a well encoded memory is retrieved in response to a closed question, and it activates other strongly encoded memories, the witness may not provide this information as it has not been requested.

The associative nature of memory can also help explain why monitoring was more effective during open-ended than closed questioning. Specifically, participants may have been better able to identify correct memories retrieved during open-ended questioning because the memories that came to mind were their most strongly encoded memories about the stimulus video. In contrast, the memories that came to mind during closed questioning may have been a mixture of memories of various strengths because the questions would have varied in difficulty. For example, there are likely to have been some questions that asked about items that were well encoded and others that asked about items that were not well encoded. This variability in question difficulty may have interfered with the participants' ability to judge accuracy and alter their response criterion accordingly.[49] In

[49] Despite the large amount of data available, an exploration of item difficulty was beyond the scope of this thesis. The only means of assessing item difficulty would have been to produce aggregate scores to identify which questions were answered correctly and incorrectly by the majority of participants; a method that does not account for individual differences. A question answered correctly by five participants would be classed as a difficult according to the aggregate score. However, the participants who answered it correctly may have been paying close attention to the detail in question and, consequently, would found the question easy to answer. Future research could consider using manipulation of item difficulty (e.g., exposure duration or picture quality) rather than relying of measurement and aggregate scores.

fact, Koriat and Goldsmith (1996) have shown that use of the *don't know* response is more effective when confidence is polarised (i.e., when items elicit either very high confidence or very low confidence) than when confidence is spread across the entire confidence scale. Furthermore, as mentioned in the introduction of this chapter and in Chapter 4, people can find it hard to adjust their metacognitive judgments to account for changes in task difficulty (Suantak et al., 1996). Thus, the open question accuracy advantage may be a consequence of more effective monitoring during open-ended questioning which results from the activation of strongly encoded memories which are easier to identify as correct.

Based on the available data, it is unclear whether learning more about the regulation strategies that underlie responses to open-ended questions will help develop ways of improving responding to closed questions. In the matched data set, accuracy did not differ significantly between open-ended and closed questioning, suggesting that the superior monitoring observed during open-ended questioning did not help participants provide a more accurate memory report. Conversely, the results for the full data set suggested the opposite conclusion because accuracy was higher for open-ended than closed questioning due to increased retrieval of correct information and more effective monitoring. The methodology of Experiment 5 made it difficult to determine which of these conclusions was the most appropriate. In the matched data set, participants had more opportunities to increase accuracy in response to closed questions because more information was retrieved for closed than open-ended questions in this data set. It is possible that the order in which the recall tasks were completed contributed to this result because research suggests that eyewitnesses can sometimes retrieve additional correct information across a series of recall attempts (Bornstein, Liebel, & Scarberry, 1998; Dunning & Stern, 1992; Eugenio, Buckhout, Kostes, & Ellison, 1982; Scrivner & Safer, 1988). Thus, the additional information retrieved in response to closed questions in the

matched data set may be due to the fact that the closed questions were always presented after the open-ended questions. If this is the case, the accuracy level observed for closed questions in the matched data set may be artificially inflated. In the full data set, the opposite problem is apparent because participants had fewer opportunities to retrieve information for closed questions as the closed questions did not assess every aspect of the video. It will be important for future studies to manipulate question type between-subjects and include a broader set of closed questions so that a similar amount of retrieval opportunities are available across the question types. This will allow for a more precise assessment of the contribution monitoring and control make to accuracy when open-ended and closed questions are used. If additional studies do find that accuracy is higher for open-ended than closed questions in these circumstances, learning more about the regulation strategies that underlie responses to open-ended questions may help develop ways of improving responding to closed questions.

In conclusion, Experiment 5 suggested that the low withholding rates observed throughout this thesis were not a consequence of the closed question that were used. However, it was difficult to compare response bias between open-ended and closed questions in this experiment overall because of the different retrieval opportunities afforded by the two question formats. Despite this, it appears that monitoring ability is superior when open-ended questioning is used. The task for future research will be to determine more precisely the contribution that this superior monitoring makes to accuracy by including a larger sample of closed questions and manipulating question type between-subjects. This will help determine whether the regulation strategies employed during open-ended questioning can assist in improving the way witnesses respond to closed questions.

CHAPTER 7 – GENERAL DISCUSSION

This thesis has described a series of studies designed to investigate ways of improving monitoring ability, with the aim of helping witnesses simultaneously maximise the quantity and accuracy of their memory reports. Based on the source monitoring framework and previous research, I proposed that informing people of differences in the mnemonic cues associated with correct and incorrect memories may be able to improve monitoring because these mnemonic cues may not be used to their full extent when people judge the accuracy of their memories. I began by examining whether the mnemonic cues identified in the source monitoring framework could predict whether memories were correct or incorrect in a situation where misinformation was not provided. I then attempted to manipulate witness knowledge of these mnemonic cues in three experiments. The results of these studies suggested that most of the mnemonic cues identified in the source monitoring framework do reliably predict response accuracy, and that some of them are not fully utilised by witnesses during monitoring. However, manipulating knowledge of one of the mnemonic cues that are not fully utilised by witnesses during monitoring did not result in better monitoring or impact on quantity or accuracy. Furthermore, warning witnesses about the fallibility of eyewitness memory in addition to manipulating mnemonic cue knowledge and providing retrieval instructions did not affect monitoring, quantity, or accuracy. As there was some indication that withholding may be inhibited by closed questions, which could result in lax monitoring, a final experiment examined monitoring and control during open-ended and closed questions. The results indicated that the low withholding rates observed in the previous experiments were not a consequence of the use of closed questions. However, there was some indication that more effective regulation strategies are employed during open-ended than closed questioning, though additional research is required to explore these effects further.

Mnemonic Cues, Response Accuracy, and Eyewitness Confidence

The first question addressed in this thesis was whether the mnemonic cues outlined in the source monitoring framework (Johnson et al., 1993) can discriminate between naturally occurring correct and incorrect memories. This question was primarily investigated in Study 1, though Experiments 2-4 helped establish the reliability of the relationship between response accuracy and a selection of the mnemonic cues. Study 1 provided evidence that both the sensory characteristics (i.e., visual detail, clarity, and vagueness) and cognitive processes (i.e., reasoning, thoughts, and effortfulness) outlined in the source monitoring framework do discriminate between correct and incorrect memories in a complex eyewitness memory task where misinformation is not provided. However, the reasoning cue was not found to be a reliable predictor of response accuracy across studies as it was unrelated to response accuracy in Experiments 2 and 3. These results demonstrate that the majority of the cues that discriminate between externally and internally derived memories described in the source monitoring framework also discriminate between naturally occurring correct and incorrect memories.

The findings also suggested that witnesses may not fully utilise all of the mnemonic cues that discriminate between correct and incorrect memories when they monitor the accuracy of their memories. Study 1 revealed that five of the seven mnemonic cues measured were significant predictors of response accuracy after confidence was taken into account. Thus, it initially appeared that witnesses do not spontaneously consider reasoning, thoughts, clarity, vagueness, or perceived retrieval fluency during monitoring, but that they do consider effortfulness and visual detail. Consistent with this, Experiment 2 suggested that the majority of people are aware that it is important to consider visual detail when providing an eyewitness memory report, and that many do not consider thoughts or

reasoning unless instructed to.[50] However, some of the other findings from Study 1 were contradicted by the results of subsequent experiments. Experiment 2 suggested that the majority of people are aware of the importance of considering clarity when providing an eyewitness memory report, while Experiments 2 and 3 suggested that most people are not fully aware of the importance of considering effortfulness unless they are told to consider this mnemonic cue. As explained in Chapter 4, these discrepancies in the results may be due to differences in the amount and type of mnemonic cues being rated across the studies. In Study 1, rating seven mnemonic cues may have meant that particular ratings were informed by others. Thus, when only one or two ratings were made in Experiments 2-4, participants may have been unable to adequately judge the mnemonic cue or cues. Similarly, some of the mnemonic cue ratings may have detracted from others, meaning that the ability of particular mnemonic cues to predict response accuracy could have been underestimated in Study 1. In sum, it appears that reasoning, thoughts, vagueness, perceived retrieval fluency, and effortfulness (but not clarity or visual detail) are not fully considered by witnesses during monitoring. However, given the discrepancies between the results, it will be important to replicate these findings in future research.

These findings also have implications for theories of confidence within eyewitness testimony. To date, research into the basis of confidence judgments in eyewitness testimony is limited, as explained in Chapter 3. However, Robinson and colleagues (1997, 2000) have found evidence that confidence judgments may be based on vividness, the extent to which a memory is reconstructed/visualised, perceived retrieval effort, and perceived retrieval fluency. In accordance with this, the results presented in this thesis suggested that visual detail and clarity are spontaneously considered by witnesses when

[50] The fact that witnesses do not appear to consider reasoning may be advantageous given that this mnemonic cue was not found to be a reliable predictor of response accuracy across my studies.

they monitor the accuracy of their memories, as explained above. However, Study 1 seemed to contest Robinson et al.'s (2000) finding that confidence judgments are based on perceived retrieval fluency, though this can probably be explained by differences in the way participants made their estimates of retrieval fluency. Specifically, providing an estimate of the number of seconds taken to answer each question may be a better cue to base confidence judgments on than a more general rating of how quickly a memory came to mind. The findings presented in this thesis also suggested that perceived retrieval effort may not be fully captured within confidence judgments. Thus without instructions, people may not fully utilise effortfulness during monitoring.

Mnemonic Cue Information, Memory Inaccuracy Warnings, and Retrieval Instructions as Methods of Improving Eyewitness Monitoring

The second, though primary, question addressed in this thesis was whether witnesses can be trained to engage in better monitoring of their own response accuracy so that they can maximise the quantity and accuracy of their eyewitness memory reports. Experiments 2-4 examined whether information about particular mnemonic cues, a warning regarding the fallibility of eyewitness memory, or instructions regarding the process of retrieving information from memory could improve witness' ability to discriminate between correct and incorrect memories. The findings suggested that monitoring cannot be improved by any of these means and that these techniques do not impact upon the quantity or accuracy of eyewitness memory reports.

The fact that providing information about mnemonic cues was unable to improve monitoring is somewhat at odds with the literature on the DRM and misinformation paradigms. As explained in Chapter 1, information about mnemonic cues that are associated with studied words in the DRM paradigm have been found to reduce the DRM

false recognition effect (Lane et al., 2008). Similarly, information about mnemonic cues that are associated with real and suggested memories have been found to reduce the misinformation effect (Bulevich & Thomas, 2012; Lane et al., 2007). However, I found no evidence that providing information about a mnemonic cue that is associated with naturally occurring correct and incorrect memories (i.e., effortfulness) is able to improve the way witnesses respond to closed questions. This may be explained by the fact that the majority of the participants in Experiments 2-4 were young adults who completed cued-recall tests. In a misinformation study, Bulevich and Thomas (Experiment 2, 2012) encouraged participants of different ages to consider visual imagery and auditory and contextual information (which they termed *supportive instructions*) when completing either a recognition or cued-recall test. For the older adults, Bulevich and Thomas (2012) observed that the supportive instructions improved resolution and accuracy amongst older participants regardless of the type of test they completed. However, the supportive instructions only improved resolution and accuracy in recognition tests for the young adults. Thus, it seems that mnemonic cue instructions may not be helpful for young adult witnesses when they respond to cued-recall questions. Bulevich and Thomas (2012) argued that this may be because younger adults automatically engage in deeper and more efficient memory searches when answering questions in a cued-recall task. However, my results suggest that even young witnesses do not consider all the important information that they should when searching their memory for answers to questions in cued-recall tasks. It could be that the older witnesses in Bulevich and Thomas' (2012) study were more willing to accept guidance about their monitoring processes given the baseline limitations of their memory (i.e., lower response accuracy than young adults). Thus, the mnemonic cue

instructions used in this thesis may be more beneficial for an older population.[51]

There are also several other reasons which may explain why the mnemonic cue instructions used in this thesis failed to have an impact on monitoring, quantity, and accuracy. In regards to the reasoning cue, it is likely that manipulating knowledge of this cue did not have any impact on the core outcome measures because it is not a reliable predictor of response accuracy. I would not have expected providing information about a mnemonic cue that is not diagnostic of response accuracy to improve monitoring or help people maximise the quantity and accuracy of their eyewitness memory reports. In regards to the effortfulness cue, it is possible that the impact of considering effortfulness is small and only detectable in a very large sample. For example, the presence of effortfulness may be more helpful to witnesses than the absence of it because correct memories are likely to be associated with other mnemonic cues that help witnesses recognise them as correct (e.g., clarity and visual detail). Thus, correct memories might be volunteered even when people do not consider the effortfulness of their memories. Incorrect memories, however, may only be distinguished by the presence of effortfulness. Thus, if witnesses do not fully consider effortfulness, they will be more likely to volunteer incorrect memories. Another problem arises if only a small number of memories are effortful to retrieve such as when the proportion of incorrect memories retrieved is low or when only some incorrect memories are effortful to retrieve. If only a small number of memories are effortful to retrieve, the effect of using effortfulness to guide control decisions will be very small when the eyewitness memory report is examined overall. Furthermore, it is possible that the effortfulness instructions provided in Experiments 2-4 were not detailed enough to produce an effect. Indeed, the results of the manipulation checks generally suggested that the

[51] It was not possible to test this in my experiments given that the average age of participants in my experiments around 22 years, with very few participants over the age of 40.

information participants were given about effortfulness were either only having a small impact on knowledge about this mnemonic cue or were only effective for a portion of the participants. Thus, it may be important to explore the impact of more detailed instructions in future studies, though this will initially require a more detailed assessment of exactly what constitutes effortfulness.

In addition to the aforementioned explanations, it could also be that the effortfulness instructions were ineffective because witnesses overestimate the accuracy of their memories. The results of Experiments 2-4, showed that participants had a very strong bias towards volunteering information. While I hypothesised that this tendency might have been explained by the use of closed questions, the findings of Experiment 5 refute this possibility as closed question did not result in higher levels of withholding than open-ended questions. Thus, it is more plausible that the tendency towards volunteering is indicative of an overestimation of memory accuracy in eyewitness memory tasks. Indeed, Perfect (2004) observed that while participants believed they would perform better on an eyewitness memory task than a general knowledge test, their predictions did not match actual performance. Furthermore, Simons and Chabris (2011) found that the majority of a US sample of participants believed memory operated like a video recorder, suggesting a strong belief in the accuracy of memory. Beliefs such as this may not be particularly problematic when people are questioned in an open-ended format because Experiment 5 suggested that open-ended questions result in the retrieval of a lot of correct information and people seem to be quite good at monitoring the accuracy of the information retrieved during open-ended questioning. However, it may be less appropriate to trust information that comes to mind when closed question are used because such questions can prompt the retrieval of incorrect information that would not be retrieved during open-ended questioning, as shown in Experiment 5. Thus, if witnesses rarely consider the possibility

that their memory could be wrong when they are answering closed questions, they may not make proper use of the effortfulness cue because the presence of effortfulness is associated with incorrect memories.

Despite the fact that misconception about memory accuracy may contribute to ineffective monitoring, it seems that a simple warning about the fallibility of eyewitness memory will be unable to affect a change in people's perceptions of the accuracy of their memories given that such a warning was found to be ineffective in Experiment 4. Future research could explore more effective means of altering the trust witnesses have in the accuracy of their memories when they are answering closed questions, though it is possible that such a change can only be achieved with relatively intensive (and therefore impractical) training as suggested by Lane and Karam-Zanders (2013). For example, Niedźwieńska (2004) found that an intensive 30-hour training course about autobiographical memory (including information about potential sources of memory error) was able to improve the accuracy of autobiographical memory reports for the September 11 terrorist attacks. Such a training regime is not feasible in the eyewitness context given the limited resources of police (i.e., time and staff) and the need to conduct interviews as soon as possible following the crime to increase the chances of apprehending the offenders before they have the opportunity to destroy evidence or commit further crimes.

Although the results presented in this thesis indicated that retrieval instructions may not be a useful means of improving the way witnesses respond to closed questions, further study on this technique may be warranted. As explained in Chapter 5, differences between the methodology used in Experiment 5 and the methodology used by Scoboria et al. (2014) might explain why the retrieval instructions were ineffective. It may be that demonstrating the instructions with an example and having witnesses verbally articulate their retrieval processes allows them to better understand and utilise the instructions, which may

consequently improve the way they respond to closed questions. The inclusion of an example would be reasonably easy to implement and test in both the laboratory and the field because participants and witnesses, respectively, could simply be provided with a written copy of the instructions and the example. Verbal articulation of the retrieval processes could also be implemented easily in the field and in experimental research, particularly given that this method is likely to require less time to conduct than the Cognitive Interview that is typically recommended for interviewing witnesses.

Retrieval, Monitoring, and Control Abilities during Open-Ended and Closed Questioning

The third question addressed in this thesis was whether the retrieval, monitoring, and/or control abilities of witnesses differ depending on the type of questions they are asked given that open-ended questions typically elicit more accurate reports than closed questions (Fisher, 1995; Lipton, 1977). The purpose of assessing this was to (i) determine whether the low withholding rates observed in Experiments 2-4 were the result of the closed questions that were used, (ii) examine underlying mechanism for the open question accuracy advantage, and (iii) evaluate the memory regulation processes that operate during open-ended questioning as this may be a potential avenue for improving the way witnesses respond to closed questions. As explained earlier, Experiment 5 revealed that the use of closed questions could not explain the low withholding rates observed in Experiments 2-4. Although it was difficult to appropriately compare monitoring and control abilities between open-ended and closed questions, there was some indication that monitoring is more effective during open-ended questioning and these questions also appear to prompt the retrieval of a large amount of additional information. Thus, learning more about the memory regulation processes that operate during open-ended questioning may offer a

means of helping witnesses maximise the quantity and accuracy of their eyewitness memory reports.

The results reinforce the well-established importance of asking open-ended questions when interviewing witnesses. The Cognitive Interview is comprised of various social dynamic, communication, and cognitive components and encourages free-narrative and open-ended questions as opposed to closed questions (Fisher & Geiselman, 1992), and is considered best practice in many developed countries because it has been found to increase the amount of correct information obtained from witnesses with only a small increase in the amount of incorrect information (Köhnken, Milne, Memon, & Bull, 1999; Memon, Meissner, & Fraser, 2010). It is. Consistent with this, the results of Experiment 5 showed that when all of the information retrieved during the open-ended questions was assessed, significantly more correct information was retrieved for open-ended than closed questions. Thus, open-ended questions resulted in the retrieval of a large amount of correct information that would not have been retrieved if only this set of closed questions had been asked. This is important because the effectiveness of closed questions is dependent upon the interviewer knowing exactly what to ask. If the interviewer does not ask questions about everything the witness can remember, they may miss potentially important information, information that could have been volunteered in response to an open-ended question. Furthermore, open-ended questions also appear to enable superior monitoring which also helps explain the usefulness of the Cognitive Interview. Experiment 5 showed that the superior monitoring exhibited by participants during open-ended questioning meant that most of the additional correct information they retrieved was volunteered, without increasing the amount of incorrect information that was volunteered. These findings reinforce the importance of asking open-ended questions when interviewing witnesses, in accordance with recommendations (Technical Working Group for Eyewitness

Evidence, 1999). In addition, the findings explain that open-ended questions elicit more correct information because they increase the amount of correct information retrieved and improve monitoring ability, allowing witnesses to volunteer more correct information than they would in response to closed questions while withholding most of the incorrect information.

The findings of Experiment 5 also confirmed the importance of improving eyewitness' monitoring during closed questioning as such questions may allow police to obtain additional correct information that is not retrieved during open-ended questioning. When only items that could be matched across the questions were considered in Experiment 5, the findings showed that closed questions resulted in the retrieval of significantly more correct information than open-ended questions. Thus, closed questions resulted in the retrieval of some correct information that would not have been retrieved if only open-ended questions had been asked. This is valuable from an applied perspective as it suggests that police may be able to use closed questions to obtain additional correct information about a crime that may be critical for the investigation or for legal reasons, assuming that the additional correct information is volunteered. However, Experiment 5 also showed that closed questions resulted in the retrieval of significantly more incorrect information, less effective monitoring, and a more conservative response bias than open-ended questions. Thus, witnesses may have difficulty determining that the additional information is correct, and they may not be confident enough in the additional correct information to volunteer it, meaning that the additional correct information may not be volunteered. Furthermore, monitoring difficulties could also make it more likely that additional incorrect information will be volunteered if witnesses are quite confident in the incorrect information they retrieve. Therefore, the findings demonstrate the importance of improving the way witnesses monitor the accuracy of their memory during closed

questions because such questions may be able to elicit valuable information that witnesses may not otherwise remember.

The results of Experiment 5 may also have implications for theories of confidence within eyewitness testimony. The fact that closed questioning appears to interfere with monitoring ability may suggest that people consider whether they should be able to answer the question, in addition to what their actual answer is (i.e., an answer or a *don't know* response). This will be problematic if the witness believes they should be able to answer a particular question but either cannot retrieve one, or can only retrieve one they are not confident about. For example, the witness may be asked what type of jacket the offender wore. As this is a central piece of information, they may believe they should know the answer. However, they may have been paying attention to other aspects of the event and neglected to fully encode the type of jacket worn by the offender. In this instance, the witness may lack confidence in any answer they retrieve but choose to volunteer it because they believe that it is something they should remember, devaluing their accurate confidence judgment. Thus, confidence judgments may be inflated in situations where witnesses answer closed questions. Furthermore, witnesses may base their confidence judgments on the speed with which answers come to mind when they answer any type of question (Robinson et al., 1997, 2000), regardless of whether it is open-ended or closed. This strategy may be useful during open-ended questioning in which the strongest (and therefore most likely to be accurate) memories are likely to be retrieved quite quickly. However, during closed questions, fluency of retrieval may be distorted by the fact that a retrieval cue is present. Thus, retrieval fluency may be a less reliable predictor of both confidence and accuracy in the context of closed questions. However, given that Study 1 suggested that perceived retrieval fluency is not spontaneously considered by witnesses when they judge the accuracy of their memory, this explanation may be less plausible.

The Global Informativeness Criterion as an Explanation for the Liberal Response Bias of Witnesses

As explained earlier, the strong tendency of witnesses to volunteer (i.e., liberal response bias) information could be explained by an overestimation of the accuracy of memory. However, it is possible that the extremely liberal response bias observed throughout the experiments presented in this thesis is a consequence of a preference to appear informative. Specifically, it may be that witnesses prefer to volunteer as much information as possible and have some of this information be incorrect than provide only information that they know for sure is correct. Ackerman and Goldsmith (2008) have proposed that when people answer a series of questions, they consider how much information they are volunteering overall when deciding whether to volunteer or withhold each answer. In an experiment where participants answered general knowledge questions and could decide how specific each answer was or withhold the answer entirely, Ackerman and Goldsmith (2008) found that participants sometimes provided specific answers that they were not confident about. Specifically, the findings showed that approximately 17% of the answers participants volunteered were associated with a level of confidence below their estimated response criterion (i.e., the level of confidence required to warrant volunteering an answer). Ackerman and Goldsmith (2008) concluded that there may be a global informativeness criterion that causes people to devalue non-specific information and/or causes them to avoid withholding information. Therefore, the low rate of withholding observed in the experiments presented in this thesis may be a consequence of a global informativeness criterion which makes witnesses feel compelled to volunteer, thus preventing them from giving *don't know* responses for a large number of questions.

From a practical perspective, it is easy to see why witnesses may feel pressured to avoid giving *don't know* responses for a large number of questions. Research into the

usage of different components of the Cognitive Interview has found that although police can have difficulty adhering to all of its components, they frequently use the *report everything* instruction (Dando et al., 2008, 2009; Kebbell et al., 1999) which is designed to discourage witnesses from withholding partial or incomplete information (Memon et al., 2010). However, it is possible that this instruction also discourages witnesses from withholding information they are unsure about. Furthermore, while the Cognitive Interview also recommends that witnesses be reminded not to guess, the reminder may only discourage blatant guessing. As a result, witnesses may continue to believe that the *report everything* instruction is encouraging them to volunteer all information, even when they are not entirely sure of its accuracy. Furthermore, research is mixed regarding how often police actually remind witnesses not to guess. While Dando et al. (2008) found that police self-reported using the reminder usually or almost always, Dando et al. (2009) found that it was not used frequently during real police interviews. Thus, if witnesses receive a *report everything* instruction, but not a *do not guess* instruction, they may volunteer information they are unsure about.

Of course, if witnesses do volunteer information they are unsure about, they could provide uncertainty qualifiers such as 'I think…', 'I'm not sure, but…', or 'It might have been…' which indicate to the police that they are uncertain about such information. Although such qualifiers are used more frequently by older witnesses, they have also been observed in young adult witnesses (Brimacombe, Quinton, Nance, & Garrioch, 1997). Future research could explore whether witnesses use more uncertainty qualifiers (or different uncertainty qualifiers) for incorrect answers that they volunteer than for correct answers that they volunteer to determine whether monitoring is occurring at this more specific level. If such an effect is found, it may be possible to provide police with information that can help them discriminate between information that is more likely to be

correct (or incorrect). Providing police with guidance that can help them discriminate between the correct and incorrect information they obtain from witnesses may be particularly important given the pragmatic demands of conducting interviews. A common barrier to conducting the Cognitive Interview in full is time constraints, with police reporting that they often have very heavy workloads and feel pressured by more senior officers to conduct their interviews quickly (Dando et al., 2008; Kebbell et al., 1999). Such pressures may be conveyed to witnesses, either implicitly or explicitly, and cause them to feel as though they must answer the interview questions as quickly as possible. It may be that it is quicker for witnesses to provide uncertainty qualifiers for the answers they are not confident about, rather than spend time engaging in careful monitoring and control processes. If police are aware of the qualifiers and how they relate to response accuracy, they may be able to determine which details they should trust and consequently follow-up as part of their investigation.

Limitations

A number of important limitations are evident throughout this thesis which may impact on interpretations and inform directions for future studies. First, the delay between presentation of the stimulus and the recall task was relatively short in each experiment (i.e., 10 minutes). However, research suggests that memory decays over time and that witnesses remember less information as the delay between the event and the recall task increases (Ebbesen & Rienick, 1998; Odinot & Wolters, 2006). Thus, the results observed in this thesis only pertain to recall tasks conducted shortly after the stimulus event because little time was allowed for forgetting to occur. Second, a single stimulus video and the same set of closed questions were used to assess memory throughout this thesis. Future research should aim to determine the extent to which the results generalise to other stimulus materials and, as noted earlier, it will be particularly important to use a broader set of

closed questions when comparing retrieval, monitoring, and control processes that operate during open-ended and closed questioning. Third, the filter questions included in the recall task added an additional component to the memory retrieval process. Such questions are known to reduce accuracy (Lipton, 1977) and although they were not included in the analyses, answering them may have impacted on the way people responded to closed questions. Thus, forthcoming studies should aim to utilise a set of recall questions that do not include forced-choice alternatives.

The fourth, though perhaps most critical limitation of the experiments presented in this thesis relates to the way in which the retrieval processes were decomposed to allow assessment of metacognitive monitoring and control. In each experiment, participants were required to answer all questions, make explicit decisions about which answers to volunteer and withhold, and provide ratings of metacognitive judgements (i.e., confidence and mnemonic cues). Although the procedures used were necessary to appropriately measure monitoring and control abilities, as well as quantity and accuracy, they are likely to be very different from the conditions under which people would normally answer questions about a crime. For example, a forensic investigator is unlikely to request a confidence judgement for each piece of information a witness volunteers, and, if the interview is conducted according to recommended practice, the witness will not be asked to answer all questions as guessing should be discouraged (Fisher, 1995; Technical Working Group, 1999). Recent evidence suggests that memory regulation processes can differ depending on whether eyewitness memory reports are obtained via a two-phase reporting procedure or a more naturalistic procedure (Sauer & Hope, 2015). Thus, there is a possibility that the results observed in this thesis do not accurately reflect the regulatory processes that people normally engage in.

The fifth, though related, area of concern in this thesis was the externalised free-recall technique used in Experiment 5 to assess retrieval, monitoring, and control during open-ended questioning. Typically, when people respond to open-ended questions, they engage in narrative recall and the information they provide could either represent everything they were able to retrieve or a subset of the retrieved information that they felt comfortable volunteering. Either way, they are able to engage in uninterrupted recall. It is possible that the two-column method used in Experiment 5 interfered with this recall process because people were required to switch columns when they retrieved something that they wanted to withhold. Interrupting recall in this way may have prevented additional information from being retrieved because it could have hindered the associative process involved in memory retrieval. Thus, Experiment 5 may have underestimated the amount of information that could be retrieved in response to the open-ended questions. However, it is important to remember that this was the first time this technique was utilised in an eyewitness recall task, and that such as task was necessary because typical procedures do not allow assessment of monitoring and control abilities. It may be useful to compare the typical recall method with a variety of the externalised free-recall methods (e.g., the asterisk method) to determine which methodology is most suited to assessing monitoring and control in eyewitness recall tasks.

General Conclusions

The experiments reported in this thesis demonstrated that the mnemonic cues outlined in the source monitoring framework can discriminate between correct and incorrect memories, though some of them are spontaneously considered by witnesses during the monitoring process. The findings also suggested that eyewitness monitoring cannot be improved by (i) providing witnesses with information about mnemonic cues that they do not take into full consideration during monitoring, (ii) warning witnesses of

fallibility of eyewitness memory, or (iii) providing retrieval instructions. However, this does not mean that it is impossible to improve monitoring, and it seems that exploring eyewitness monitoring during open-ended questions may help uncover ways of improving the way people respond to closed questions. Two important avenues for future research will be to (i) use a broader set of closed questions to enable a more rigorous comparison of the monitoring and control processes during open-ended and closed questioning, and (ii) uncover why witnesses exhibit such a liberal response bias. Forthcoming experiments could explore whether the liberal response bias of witnesses is the result of an overestimation of the accuracy of memory and/or a global informativeness criterion. In addition, it may be important to explore whether witnesses provide other information, perhaps in the form of uncertainty qualifiers, which can give police some indication of whether the information they volunteer is correct.

REFERENCES

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

Adams, C., Smith, M. C., Pasupathi, M., & Vitolo, L. (2002). Social context effects on story recall in older and younger women: Does the listener make a difference? *The Journals of Gerontology: Series B: Psychological Sciences and Social Sciences*, *57B*, 28–40. http://doi.org/10.1093/geronb/57.1.P28

Anastasi, J. S., Rhodes, M. G., & Burns, M. C. (2000). Distinguishing between memory illusions and actual memories using phenomenological measurements and explicit warnings. *The American Journal of Psychology*, *113*, 1–26. http://doi.org/10.2307/1423458

Anderson, J. R. (1976). *Language, memory and thought*. Hillsdale: Erlbaum Associates

Anderson, J. R. (1983a). *The architecture of cognition*. Cambridge: Harvard University Press

Anderson, J. R. (1983b). A spreading activation theory of memory. *Journal of Verbal Learning and Verbal Behaviour*, *22*, 261–295. http://dx.doi.org/10.1016/S0022-5371(83)90201-3

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390–412. http://doi.org/10.1016/j.jml.2007.12.005

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*, 255–278. http://doi.org/10.1016/j.jml.2012.11.001

Bates, D. M. (2007). Computational methods for mixed models [Tech. rep]. Madison, WI: Dept. of Statistics, University of Wisconsin.

Bates, D. M., Maechler, M., Bolker, B., & Walker, S. (2015). _lme4: Linear mixed-effects models using Eigen and S4_. Retrieved from http://CRAN.R-project.org/package=lme4

Benson, P. G., & Önkal, D. (1992). The effects of feedback and training on the performance of probability forecasters. *International Journal of Forecasting*, *8*, 559–573. http://doi.org/10.1016/0169-2070(92)90066-i

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*, 77–88. http://doi.org/10.1016/j.jarmac.2014.03.005

Bornstein, B. H., Liebel, L. M., & Scarberry, N. C. (1998). Repeated testing in eyewitness memory: A mean to improve recall of a negative emotional event. *Applied Cognitive Psychology*. *12*, 119–131. http://dx.doi.org/10.1002/(SICI)1099-0720(199804)12:2%3C119::AID-ACP500%3E3.0.CO;2-4

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*, 76–88. http://doi.org/10.1037//1076-898X.5.1.76

Brigham, J. C., & Cairns, D. (1988). The effect of mugshot inspections on eyewitness identification accuracy. *Journal of Applied Social Psychology*, *18*, 1393–1410. http://doi.org/10.1111/j.1559-1816.1988.tb01214.x

Brimacombe, C. A. E., Quinton, N., Nance, N., & Garrioch, L. (1997). Is age irrelevant? Perceptions of young and old adult eyewitnesses. *Law and Human Behavior*, *21*, 619–634. http://doi.org/10.1023/A:1024808730667

Bulevich, J. B., & Thomas, A. K. (2012). Retrieval effort improves memory and metamemory in the face of misinformation. *Journal of Memory and Language*, *67*, 45–58. http://doi.org/10.1016/j.jml.2011.12.012

Carneiro, P., & Fernandez, A. (2013). Retrieval dynamics in false recall: revelations from identifiability manipulations. *Psychonomic Bulletin & Review*, *20*, 488–495. http://doi.org/10.3758/s13423-012-0361-4

Christiaansen, R. E., & Ochalek, K. (1983). Editing misleading information from memory: Evidence for the coexistence of original and postevent information. *Memory & Cognition*, *11*, 467–475. http://doi.org/10.3758/BF03196983

Craik, F. I., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology*, *104*, 268–294. http://dx.doi.org/10.1037/0096-3445.104.3.268

Dando, C., Wilcock, R., & Milne, R. (2008). The cognitive interview: Inexperienced police officers' perceptions of their witness/victim interviewing practices. *Legal and Criminological Psychology*, *13*, 59. http://doi.org/10.1348/135532506X162498

Dando, C., Wilcock, R., & Milne, R. (2009). The cognitive interview: Novice police officers' witness/victim interviewing practices. *Psychology, Crime & Law*, *15*, 679–696. http://doi.org/10.1080/10683160802203963

Dixon, P. (2008). Models of accuracy in repeated-measures designs. *Journal of Memory and Language*, *59*, 447–456. http://doi.org/10.1016/j.jml.2007.11.004

Dunlosky, J., & Metcalfe, J. (2009). *Metacognition*. Thousand Oaks: Sage Publications.

Dysart, J. E., Lindsay, R. C. L., Hammond, R., & Dupuis, P. (2001). Mugshot exposure prior to lineup identification: Interference, and commitment effects. *Journal of Applied Psychology, 86,* 1280–1284. http://dx.doi.org/10.1037/0021-9010.86.6.1280

Ebbesen, E. B., & Rienick, C. B. (1998). Retention interval and eyewitness memory for events and personal identifying attributes. *Journal of Applied Psychology, 83,* 745–762. http://dx.doi.org/10.1037/0021-9010.83.5.745

Echterhoff, G., Groll, S., & Hirst, W. (2007). Tainted truth: Overcorrection for misinformation influence on eyewitness memory. *Social Cognition, 25*, 367–409. http://doi.org/10.1521/soco.2007.25.3.367

Ericsson, K., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review, 87,* 215–251. http://doi.org/10.1037//0033-295X.87.3.215

Eugenio, P., Buckhout, T., Kostes, S. & Ellison, K. (1982). Hypermnesia in the eyewitness to a crime. *Bulletin of the Psychonomic Society, 19*, 83–86. http://dx.doi.org/10.3758/BF03330047

Fisher, R. P. (1995). Interviewing victims and witnesses of crime. *Psychology, Public Policy, & Law Special Theme: Witness Memory and Law*, *1*, 732–764. http://doi.org/10.1037/1076-8971.1.4.732

Fisher, R. P., & Geiselman, R. E. (1992). *Memory enhancing techniques for investigative interviewing: The Cognitive Interview*. Springfield, IL: Charles C. Thomas.

Gallo, D. A., Roberts, M. J., & Seamon, J. G. (1997). Remembering words not presented in lists: Can we avoid creating false memories? *Psychonomic Bulletin & Review, 4*, 271–276. http://doi.org/10.3758/BF03209405

Gallo, D. A., Roediger, H. L., & McDermott, K. B. (2001). Associative false recognition occurs without strategic criterion shifts. *Psychonomic Bulletin & Review*, *8*, 579–586. http://doi.org/10.3758/BF03196194

Garson, G. D. (2012). Fundamentals of hierarchical linear and multilevel modeling. In G. D. Garson (Ed.), *Hierarchical linear modeling: Guide and applications*. Thousand Oaks, USA: Sage Publications.

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge, New York: Cambridge University Press.

Goldsmith, M., & Koriat, A. (2008). The strategic regulation of memory accuracy and informativeness. In A. S. Benjamin & B. H. Ross (Eds.), *Skill and strategy in memory use* (pp. 1–60). San Diego, CA, US: Elsevier Academic Press.

Goodsell, C. A., Gronlund, S. D., & Neuschatz, J. S. (2015). Investigating mug shot commitment. *Psychology, Crime & Law*, *21*, 219–233. http://doi.org/10.1080/1068316X.2014.951647

Gorenstein, G. W., & Ellsworth, P. C. (1980). Effect of choosing an incorrect photograph on a later identification by an eyewitness. *Journal of Applied Psychology*, *65*, 616–622. http://dx.doi.org/10.1037/0021-9010.65.5.616

Greene, E., Flynn, M. S., & Loftus, E. F. (1982). Inducing resistance to misleading information. *Journal of Verbal Learning and Verbal Behavior*, *21*, 207–219. http://doi.org/10.1016/S0022-5371(82)90571-0

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics 3: Speech acts* (pp. 41–58). New York: Academic Press.

Higham, P. A. (2002). Strong cues are not necessarily weak: Thomson and Tulving (1970) and the encoding specificity principle revisited. *Memory & Cognition*, *30*, 67–80. http://doi.org/10.3758/BF03195266

Higham, P. A., Luna, K., & Bloomfield, J. (2011). Trace-strength and source-monitoring accounts of accuracy and metacognitive resolution in the misinformation paradigm. *Applied Cognitive Psychology*, *25*, 324–335. http://doi.org/10.1002/acp.1694

Higham, P. A., & Tam, H. (2005). Generation failure: Estimating metacognition in cued recall. *Journal of Memory and Language*, *52*, 595–617. http://doi.org/10.1016/j.jml.2005.01.015

Hollins, T. J., Lange, N., Berry, C. J., & Dennis, I. (under review). The role of early selection and late correction processes in source-based recall. *Journal of Memory and Language*.

Hollins, T. J., Lange, N., Dennis, I., & Longmore, C. A. (2015). Social influences on unconscious plagiarism and anti-plagiarism. *Memory*, *0*, 1–19. http://doi.org/10.1080/09658211.2015.1059857

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*, 434–446. http://doi.org/10.1016/j.jml.2007.11.007

Johnson, M. K., Foley, M. A., Suengas, A. G., & Raye, C. L. (1988). Phenomenal characteristics of memories for perceived and imagined autobiographical events. *Journal of Experimental Psychology: General*, *117*, 371–376. http://doi.org/10.1037//0096-3445.117.4.371

Johnson, M. K., Hashtroudi, S., & Lindsay, S. D. (1993). Source Monitoring. *Psychological Bulletin*, *114*, 3–28. http://doi.org/10.1037//0033-2909.114.1.3

Johnson, M. K., Kahan, T. L., & Raye, C. L. (1984). Dreams and reality monitoring. *Journal of Experimental Psychology: General*, *113*, 329–344. http://doi.org/10.1037//0096-3445.113.3.329

Johnson, M. K., & Raye, C. L. (1981). Reality monitoring. *Psychological Review*, *88*, 67–85. http://doi.org/10.1037//0033-295X.88.1.67

Jou, J., & Foreman, J. (2007). Transfer of learning in avoiding false memory: The roles of warning, immediate feedback, and incentive. *The Quarterly Journal of Experimental Psychology*, *60*, 877–896. http://doi.org/10.1080/17470210600831184

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality & Social Psychology*, *103*, 54–69. http://doi.org/10.1037/a0028347

Kahana, M. J., Dolan, E. D., Sauder, C. L., & Wingfield, A. (2005). Intrusions in episodic recall: age differences in editing of overt responses. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *60*, P92–P97. http://doi.org/10.1093/geronb/60.2.P92

Kebbell, M. R., & Milne, R. (1998). Police officers' perceptions of eyewitness performance in forensic investigations. *The Journal of Social Psychology*, *138*, 323–330. http://doi.org/10.1080/00224549809600384

Kebbell, M. R., Milne, R., & Wagstaff, G. F. (1999). The cognitive interview: A survey of its forensic effectiveness. *Psychology, Crime & Law*, *5*, 101–115. http://doi.org/10.1080/10683169908414996

Kliegl, R., Wei, P., Dambacher, M., Yan, M., & Zhou, X. (2011). Experimental effects and individual differences in linear mixed models: Estimating the relation of spatial, object, and attraction effects in visual attention. *Frontiers in Psychology*, *1*. http://doi.org/10.3389/fpsyg.2010.00238

Köhnken, G., Milne, R., Memon, A., & Bull, R. (1999). The cognitive interview: A meta-analysis. *Psychology, Crime & Law*, *5*, 3–27. http://doi.org/10.1080/10683169908414991

Koriat, A., & Goldsmith, M. (1994). Memory in naturalistic and laboratory contexts: Distinguishing the accuracy-oriented and quantity-oriented approaches to memory assessment. *Journal of Experimental Psychology: General*, *123*, 297–315. http://doi.org/10.1037//0096-3445.123.3.297

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

Lampinen, J. M., Neuschatz, J. S., & Payne, D. G. (1997). Memory illusions and consciousness: Examining the phenomenology of true and false memories. *Current Psychology*, *16*, 181–224. http://doi.org/10.1007/s12144-997-1000-5

Lane, S. M., & Karam-Zanders, T. (2013). What do lay people believe about memory?. In T. J. Perfect & D. S. Lindsay (Eds.), *The SAGE Handbook of Applied Memory* (pp. 348–365. London: SAGE Publications

Lane, S. M., Roussel, C. C., Starns, J. J., Villa, D., & Alonzo, J. D. (2008). Providing information about diagnostic features at retrieval reduces false recognition. *Memory*, *16*, 836–851. http://doi.org/10.1080/0965821080233734

Lane, S. M., Roussel, C. C., Villa, D., & Morita, S. K. (2007). Features and feedback: Enhancing metamnemonic knowledge at retrieval reduces source-monitoring errors. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *33*, 1131–1142. http://doi.org/10.1037/0278-7393.33.6.1131

Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? *Organizational Behavior & Human Performance*, *20*, 159–183. http://doi.org/10.1016/0030-5073%2877%2990001-0

Lipton, J. P. (1977). On the psychology of eyewitness testimony. *Journal of Applied Psychology*, *62*, 90–95. http://doi.org/10.1037/0021-9010.62.1.90

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

Mazzoni, G. (2002). Naturally occurring and suggestion-dependent memory distortions: The convergence of disparate research traditions. *European Psychologist*, *7*, 17–30. http://doi.org/10.1027//1016-9040.7.1.17

McCabe, D. P., & Smith, A. D. (2002). The effect of warnings on false memories in young and older adults. *Memory & Cognition*, *30*, 1065–1077. http://doi.org/10.3758/BF03194324

McCabe, D. P., & Soderstrom, N. C. (2011). Recollection-based prospective metamemory judgments are more accurate than those based on confidence: Judgments of remembering and knowing (JORKs). *Journal of Experimental Psychology: General*, *140*, 605–621. http://doi.org/10.1037/a0024014

McDermott, K. B., & Roediger, H. L. (1998). Attempting to Avoid Illusory Memories: Robust False Recognition of Associates Persists under Conditions of Explicit Warnings and Immediate Testing. *Journal of Memory and Language*, *39*, 508–520. http://doi.org/10.1006/jmla.1998.2582

Memon, A., Hope, L., Bartlett, J., & Bull, R. (2002). Eyewitness recognition errors: The effects of mugshot viewing and choosing in young and old adults. *Memory & Cognition*, *30*, 1219–1227. http://doi.org/10.3758/BF03213404

Memon, A., Meissner, C. A., & Fraser, J. (2010). The cognitive interview: A meta-analytic review and study space analysis of the past 25 years. *Psychology, Public Policy, and Law*, *16*, 340–372. http://doi.org/10.1037/a0020518

Miller, M. B., Guerin, S. A., & Wolford, G. L. (2011). The strategic nature of false recognition in the DRM paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*, 1228–1235. http://doi.org/10.1037/a0024539

Mueller, M. L, Tauber, S. K., & Dunlosky, J. (2012). Contributions of beliefs and processing fluency to the effect of relatedness on judgments of learning. *Psychonomic Bulletin & Review*, *20*, 387–384. http://doi.org/10.3758/s13423-012-0343-6

Multhaup, K. S., & Conner, C. A. (2002). The effects of considering nonlist sources on the Deese–Roediger–McDermott memory illusion. *Journal of Memory and Language*, *47*, 214–228. http://doi.org/10.1016/s0749-596x(02)00007-4

Murayama, K., Sakaki, M., Yan, V. X., & Smith, G. M. (2014). Type I error inflation in the traditional by-participant analysis to metamemory accuracy: A generalized mixed-effects model perspective. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *40*. http://doi.org/10.1037/a0036914

Neuschatz, J. S., Benoit, G. E., & Payne, D. G. (2003). Effective warnings in the Deese-Roediger-McDermott false-memory paradigm: The role of identifiability. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *29*, 35–41. http://doi.org/10.1037/0278-7393.29.1.35

Neuschatz, J. S., Payne, D. G., Lampinen, J. M., & Toglia, M. P. (2001). Assessing the effectiveness of warnings and the phenomenological characteristics of false memories. *Memory*, *9*, 53–71. http://doi.org/10.1080/09658210042000076

Niedźwieńska, A. (2004). Metamemory knowledge and the accuracy of flashbulb memories. *Memory*, *12*, 603–613. http://doi.org/10.1080/09658210344000134

Norman, K. A., & Schacter, D. L. (1997). False recognition in younger and older adults: Exploring the characteristics of illusory memories. *Memory & Cognition*, *25*, 838–848. http://doi.org/10.3758/bf03211328

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

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

Perfect, T. J. (2004). The role of self-rated ability in the accuracy of confidence judgements in eyewitness memory and general knowledge. *Applied Cognitive Psychology*, *18*, 157–168. http://doi.org/10.1002/acp.952

Powell, M. B., Fisher, R. P., & Wright, R. (2005). Investigative interviewing. In N. Brewer & K. D. Williams (Eds.), *Psychology and Law: An Empirical Perspective* (pp. 11–42). New York: Guilford Press.

Quené, H., & van den Bergh, H. (2008). Examples of mixed-effects modeling with crossed random effects and with binomial data. *Journal of Memory and Language*, *59*, 413–425. http://doi.org/10.1016/j.jml.2008.02.002

R Development Core Team. (2015). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Robinson, M. D., Johnson, J. T., & Herndon, F. (1997). Reaction time and assessments of cognitive effort as predictors of eyewitness memory accuracy and confidence.

Journal of Applied Psychology, *82*, 416–425. http://doi.org/10.1037/0021-9010.82.3.416

Robinson, M. D., Johnson, J. T., & Robertson, D. A. (2000). Process versus content in eyewitness metamemory monitoring. *Journal of Experimental Psychology: Applied*, *6*, 207–221. http://doi.org/10.1037/1076-898X.6.3.207

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *21*, 803–814. http://doi.org/10.1037//0278-7393.21.4.803

Sauer, J. D., & Hope, L. (2015, June). The effects of divided attention at study and reporting procedure on monitoring and grain size regulation for cued recall. Paper presented at the meeting of the XIth Society for Applied Research in Memory and Cognition Conference, Victoria, Canada. Abstract retrieved from http://static1.squarespace.com/static/504170d6e4b0b97fe5a59760/t/55887e52e4b0eb99d4b5b905/1435008594892/Program15June.pdf

Schooler, J. W., Gerhard, D., & Loftus, E. F. (1986). Qualities of the unreal. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *12*, 171–181. http://doi.org/10.1037//0278-7393.12.2.171

Scoboria, A., Memon, A., Trang, H., & Frey, M. (2014). Improving responding to questioning using a brief retrieval training. *Journal of Applied Research in Memory and Cognition*, *2*, 210–215. http://doi.org/10.1016/j.jarmac.2013.09.001

Scrivner, E., & Safer, M. A. (1988). Eyewitnesses show hypermnesia for details about a violent event. *Journal of Applied Psychology, 73*, 371–377. http://dx.doi.org/10.1037/0021-9010.73.3.371

Sharp, G. L., Cutler, B. L., & Penrod, S. D. (1988). Performance feedback improves the resolution of confidence judgments. *Organizational Behavior and Human Decision Processes, 42,* 271–283. http://doi.org/10.1016/0749-5978(88)90001-5

Simons, D. J., & Chabris, C. F. (2001). What people believe about how memory works: A representative survey of the U.S. population. *PLos One, 8*. http://doi.org/http://dx.doi.org.ezproxy.flinders.edu.au/10.1371/journal.pone.0022757

Starns, J. J., Lane, S. M., Alonzo, J. D., & Roussel, C. C. (2007). Metamnemonic control over the discriminability of memory evidence: A signal detection analysis of warning effects in the associative list paradigm. *Journal of Memory and Language, 56,* 592–607. http://doi.org/10.1016/j.jml.2006.08.013

Stone, E. R., & Opel, R. B. (2000). Training to improve calibration and discrimination: The effects of performance and environmental feedback. *Organizational Behavior and Human Decision Processes, 83,* 282–309. http://doi.org/10.1006/obhd.2000.2910

Suantak, L., Bolger, F., & Ferrell, W. R. (1996). The hard–easy effect in subjective probability calibration. *Organizational Behavior and Human Decision Processes, 67,* 201–221. http://doi.org/10.1006/obhd.1996.0074

Szpitalak, M., & Polczyk, R. (2010). Warning against warnings: Alerted subjects may perform worse. Misinformation, involvement and warning as determinants of witness testimony. *Polish Psychological Bulletin*, *41*, 105. http://doi.org/http://dx.doi.org/10.2478/v10059-010-0014-2

Technical Working Group for Eyewitness Evidence. (1999). *Eyewitness evidence: A guide for law enforcement* (No. NCJ 178240). Retrieved from US Department of Justice,

Office of Justice Program, National Institute of Justice website: https://www.ncjrs.gov/pdffiles1/nij/178240.pdf

Thomas, S. J. (2004). *Using web and paper questionnaires for data-based decision making: Form design to interpretations of the results*. Thousand Oaks: Corwin Press.

Tousignant, J. P., Hall, D., & Loftus, E. F. (1986). Discrepancy detection and vulnerability to misleading postevent information. *Memory & Cognition*, *14*, 329–338. http://doi.org/10.3758/BF03202511

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

Unsworth, N., Brewer, G. A., & Spillers, G. J. (2010). Understanding the dynamics of correct and error responses in free recall: Evidence from externalized free recall. *Memory & Cognition*, *38*, 419–30. http://doi.org/10.3758/MC.38.4.419

Watson, J. M., McDermott, K. B., & Balota, D. A. (2004). Attempting to avoid false memories in the Deese/Roediger--McDermott paradigm: Assessing the combined influence of practice and warnings in young and old adults. *Memory & Cognition*, *32*, 135–41. http://doi.org/10.3758/BF03195826

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

Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology*, *143*, 2020–2045. http://doi.org/10.1037/xge0000014

Williamson, P., Weber, N., & Timmins, S. (2012). The role of intuitive statistical knowledge in confidence-accuracy calibration: How people make confidence judgments when guessing. In A. M. Columbus (Ed.), *Advances in Psychology Research* (Vol. 95, pp. 27–50). Hauppauge NY: Nova Science Publishers.

Winter, B. (2013). Linear models and linear mixed effects models in R with linguistic applications. Retrieved from http://arxiv.org/ftp/arxiv/papers/1308/1308.5499.pdf

Wright, A. M., & Alison, L. (2004). Questioning sequences in Canadian police interviews: Constructing and confirming the course of events? *Psychology, Crime & Law*, *10*, 137–154. http://doi.org/10.1080/1068316031000099120

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

APPENDIX A: CLOSED AND FILTER QUESTIONS USED IN RECALL TASKS ACROSS ALL STUDIES/EXPERIMENTS

1. What was the sex of the robber by the counter?
2. What was the general build of the robber by the counter?
3. Was the robber by the counter wearing jacket?
    a. What colour was the jacket?
    b. What type of jacket was it?
4. Was the robber by the counter wearing a shirt?
    a. What colour was the shirt?
    b. What type of shirt was it?
5. Was the robber by the counter wearing a jumper?
    a. What colour was the jumper?
    b. What type of jumper was it?
6. What colour were the trousers of the robber by the counter?
7. What type of trousers did the robber by the counter have on?
8. Was the robber by the counter wearing a disguise?
    a. What colour was the disguise?
    b. What type of disguise was it?
9. Did the robber by the counter have a weapon?
    a. What type of weapon?
10. Was the robber by the counter carrying a bag?
    a. What colour was the bag?
    b. What type of bag was it?
11. What was the sex of the robber off to the side?
12. What was the general build of the robber off to the side?

13. Was the robber by the counter wearing jacket?
    a. What colour was the jacket?
    b. What type of jacket was it?
14. Was the robber by the counter wearing a shirt?
    a. What colour was the shirt?
    b. What type of shirt was it?
15. Was the robber by the counter wearing a jumper?
    a. What colour was the jumper?
    b. What type of jumper was it?
16. What colour were the trousers of the robber off to the side?
17. What type of trousers did the robber off to the side have on?
18. Was the robber off to the side wearing a disguise?
    a. What colour was the disguise?
    b. What type of disguise was it?
19. Did the robber off to the side have a weapon?
    a. What type of weapon?
20. Was the robber off to the side carrying a bag?
    a. What colour was the bag?
    b. What type of bag was it?
21. Did the robbers use a getaway car?
    a. What colour was the getaway car?
    b. What type of car was it?
22. Did the robbers escape together?
    a. (Yes) How did they escape?
    a. (No) How did the robber that was by the counter escape?

b. (No) How did the robber that was off to the side escape?

APPENDIX B: EXPLANATION OF RESPONSE CODING FOR CLOSED QUESTIONS

Rather than coding responses by question (i.e., coding all of the responses for question one, followed by all of the responses for question two, etc.), responses were coded by participant (i.e., all of the first participant's responses were coded first, followed by all of the responses for the second participant, etc.). This was done because some participants appeared to confuse the robbers (i.e., responses for the questions about the robber by the counter were consistent with a description of the robber that was off to the side and vice versa). Had responses been coded by question, it would have been difficult to determine when participants had confused the robbers, which would have resulted in miscoding of the responses. For example, if a participant who confused the robbers said the robber by the counter was wearing a yellow hoodie, these responses (yellow and hoodie) would be coded as incorrect because the robber by the counter wore a grey suit jacket. The robber off to the side, however, did wear a yellow jumper, meaning their responses should actually be coded as correct.

There were various instances in which participants provided a response about a particular item in the wrong question (i.e., when asked for a description of the robber's disguise, they provided information about the clothing worn by the robber). However, participants typically repeated information that was given in the wrong question in the right question. When they did not repeat the response in the right question, the response was ignored for the wrong question and the appropriate code was entered into the right question. This meant that responses were not being discarded based on minor misunderstanding of the questions.

APPENDIX C: RECALL TASK INSTRUCTIONS FOR EXPERIMENT 2

Experimental condition			
Clarity + reasoning	Clarity + thoughts	Visual detail + reasoning	Confidence
You will now be asked a series of questions about the video you saw earlier. Each question will be presented on a separate screen. We would like you to imagine that you are giving a formal statement to the police. It is important that you provide as much information as possible. It is also important that the information you provide is correct. Some of the questions require a yes/no response. For these questions, the screen will show the question and 2 response buttons marked "Yes" and "No". Your task is to select the appropriate button. You will then be asked to rate how confident you are that your decision is correct. There are also some questions that require a typed response. For these questions, your task is to type an answer in the text box provided and click "Done".			
You will then be asked to complete 2 questions about your memory. The questions will be presented on separate screens. Each screen will display a statement and 5 response options ranging from "Not at All" to "Very Much". Your task is to select the option that best describes your memory.			
You will then be asked to rate how confident you are that the answer is correct. You will be presented with a confidence scale ranging from 0-100% with 10% intervals (e.g., 0%, 10%, 20% etc.). Select the percentage that corresponds to your level of confidence. Finally, you will be asked whether you want to submit the answer or withhold it from your eyewitness report. It is important that you only provide correct information. Incorrect eyewitness reports can seriously hinder an investigation and can result in the offenders escaping arrest. You should only submit answers that you think are correct. It is also important that you do not withhold correct information. Police need to have all of the correct information in order to catch the offenders. You should do your best to avoid withholding answers that you think are correct. If you want to include an answer in your eyewitness report because you think it is correct, select "Submit Answer". If you want to withhold the			

Clarity + reasoning	Clarity + thoughts	Visual detail + reasoning	Confidence
answer from your eyewitness report because you think it is incorrect, select "Withhold Answer". Remember, police need to be sure that the information you are providing is correct. Therefore, you should try to only submit correct answers. Also keep in mind that the police need all of the correct information that you know. Therefore, you should avoid withholding correct answers.			
When deciding whether an answer is correct or incorrect, we would like you consider the characteristics of your memory. Research has identified 2 characteristics that can help distinguish between correct memories and incorrect memories.			When deciding whether an answer is correct or incorrect, we would like you to consider your level of confidence. Research has found that confidence is able to distinguish between correct and incorrect memories.
It has been found that clear memories are likely to be correct. When your memory is clear, your answer is likely to be correct, and you should submit it.		It has been found that visually detailed memories are likely to be correct. When your memory is visually detailed, your answer is likely to be correct, and you should submit it.	Memories that people are highly confident about are likely to be correct. When you are highly confident in your answer, it is likely to be correct, and you should submit it.
It has also been found that memories that involve reasoning	It has also been found that memories that include a lot of	It has also been found that memories that are effortful are	Memories that people are not confident about are likely to be

Clarity + reasoning	Clarity + thoughts	Visual detail + reasoning	Confidence
are likely to be incorrect. When your memory involves reasoning, your answer is likely to be incorrect, and you should withhold it.	thoughts are likely to be incorrect. When your memory includes a lot of thoughts, your answer is likely to be incorrect, and you should withhold it.	likely to be incorrect. When your memory is effortful, your answer is likely to be incorrect, and you should withhold it.	incorrect. When you are not confident in your answer, it is likely to be incorrect, and you should withhold it.
You will now be given a yes/no practice question and a question requiring a typed response. You will be asked to type a response even if you respond "No" to the yes/no question.			
PRACTICE FILTER QUESTION: Were there any customers in the bank at the time of the robbery? PRACTICE CLOSED QUESTION: How many customers were in the bank at the time of the robbery?			
You have now completed the practice question. For the interview questions, we will not ask for a typed response when you respond "No" to a yes/no question. However, we wanted to give you practice answering a question that requires a typed response. Because there were two robbers, they will be referred to by their location during the robbery. One will be referred to as "the robber by the counter". This is the robber who took the money from the cashier. The other robber will be referred to as "the robber off to the side". There will be several questions about the top/s possibly worn by each robber. You will be asked if each robber was wearing a jacket, a shirt, and/or a jumper. Please ensure that you respond in the appropriate question. For example, if you remember that one of the robbers was wearing a jumper, but not a jacket, respond "No" to the question that asks if they were wearing a jacket, and "Yes" to the question that asks if they were wearing a jumper. If you have any questions, please ask the experimenter now.			

Clarity + reasoning	Clarity + thoughts	Visual detail + reasoning	Confidence
When deciding whether your answers are correct or incorrect, please remember to consider the characteristics of your memory.			When deciding whether your answers are correct or incorrect, please remember to consider your level of confidence.
When your memory is clear, your answer is likely to be correct, and you should submit it.		When your memory is visually detailed, your answer is likely to be correct, and you should submit it.	When you are highly confident, your answer is likely to be correct, and you should submit it.
When your memory involves reasoning, your answer is likely to be incorrect, and you should withhold it.	When your memory includes a lot of thoughts, your answer is likely to be incorrect, and you should withhold it.	When your memory is effortful, your answer is likely to be incorrect, and you should withhold it.	When you are not confident, your answer is likely to be incorrect, and you should withhold it.

APPENDIX D: RECALL TASK INSTRUCTIONS FOR EXPERIMENT 3

Experimental condition		
Reasoning	Effortfulness	Control
You will now be asked a series of questions about the video you saw earlier. Each question will be presented on a separate screen. We would like you to imagine that you are giving a formal statement to the police. It is important that you provide as much information as possible. It is also important that the information you provide is correct. Some of the questions require a yes/no response. For these questions, the screen will show the question and 2 response buttons marked Yes and No. Your task is to select the appropriate button. There are also some open-ended questions. When these are presented you should think of an answer before clicking the "Next" button.		
You will then be asked to rate the extent to which your memory involves reasoning on a 5-point scale ranging from "Not at All" to "Very Much". Your task is to select the option that best describes your memory.	You will then be asked to rate how effortful your memory was to retrieve on a 5-point scale ranging from "Not at All" to "Very Much". Your task is to select the option that best describes your memory.	
You will then be asked whether you want to provide your answer. There will be two response buttons on the screen, "Answer" and "Don't Know". To provide your answer, select the "Answer" button. To withhold your answer, select the "Don't Know" button.		
When deciding whether to provide your answer, we would like you to consider whether your memory involves reasoning.	When deciding whether to provide your answer, we would like you to consider the effortfulness of your memory. Research has	

Reasoning	Effortfulness	Control
Research has found that the level of reasoning involved in memory is able to distinguish between correct and incorrect memories. Memories that involve a lot of	found that the effort involved in retrieving information from memory is able to distinguish between correct and incorrect memories. Memories that are effortful to retrieve	
reasoning are likely to be incorrect. When your memory does involve reasoning, it is likely to be incorrect, and you should select "Don't Know".	are likely to be incorrect. When your memory is effortful to recall, it is likely to be incorrect, and you should select "Don't Know".	
It is important that you only provide correct information. Incorrect eyewitness reports can seriously hinder an investigation and can result in the offenders escaping arrest. You should only provide your answer if you think it is correct. You should say "Don't Know" if you are unsure about your answer. When you decide to provide your answer, you will be presented with another screen containing a text box in which you can type your answer. Select "Done" when you have finished typing your answer. When you select the "Don't Know" button, you will proceed to the next question. You will now be given an open-ended practice question.		
PRACTICE FILTER QUESTION: Were there any customers in the bank at the time of the robbery? PRACTICE CLOSED QUESTION: How many customers were in the bank at the time of the robbery?		
You have now completed the practice question. Please press "Next" to continue. Because there were two robbers, they will be referred to by		

Reasoning	Effortfulness	Control
their location during the robbery. One will be referred to as "the robber by the counter". This is the robber who took the money from the cashier. The other robber will be referred to as "the robber off to the side". There will be several questions about the top/s possibly worn by each robber. You will be asked if each robber was wearing a jacket, a shirt, and/or a jumper. Please ensure that you respond in the appropriate question. For example, if you remember that one of the robbers was wearing a jumper, but not a jacket, respond "No" to the question that asks if they were wearing a jumper.		
When deciding whether to provide your answer, please remember to consider whether your memory involves reasoning. When your memory involves reasoning, your memory is likely to incorrect, and you should select "Don't Know".	When deciding whether to provide your answer, please remember to consider whether your memory if effortful to recall. When your memory is effortful to recall, your memory is likely to incorrect, and you should select "Don't Know".	
RECALL TASK (see Appendix A for list of filter and closed questions)		
You have finished the interview questions. Now we would like you to think back to the questions where you responded "Don't Know". Responding "Don't Know" when you are unsure about your answer is good. However, we would like to know what your best guess was for the questions where you responded "Don't Know". You are going to be presented with these questions again. Please, think back to the answer you thought of when you were first presented with these questions. Do not try to come up with new answers. On each screen, there will be a question, a text box, and a button marked "Done". Please type the answer you were considering when you were first presented with the question and click "Done".		

APPENDIX E: RECALL TASK INSTRUCTIONS FOR EXPERIMENT 4

Experimental condition		
Training	Training + retrieval	Control
You will now be asked a series of questions about the video you saw earlier. Each question will be presented on a separate screen. We would like you to imagine that you are giving a formal statement to the police.		
It is important to understand that you will remember more incorrect information than you think. There are many factors that can cause a witness to remember incorrect information. For example, there may be issues with perception/attention because witnesses can't look at the entire event at once; also some things may be too dark to see. Witnesses may also consider related information or prior beliefs that impact on what they remember later. For these reasons and more, witnesses always remember more incorrect information than they think they do. This means that you will remember more incorrect information than you think you will. Please keep this in mind throughout the questions you will answer soon.		You should take your statement very seriously because the information you provide will almost certainly play a vital role in the investigation. Approximately 85% of police say that their major leads usually come from witnesses. This means that the statement you provide could be critical for the arrest and conviction of the perpetrators. It is absolutely essential that the information you provide is correct. Do not provide any information that could be wrong. If you provide information that is wrong, it could send the investigation off track and result in the offenders escaping arrest. You should only provide correct

Training	Training + retrieval	Control
		information.
Some of the questions require a yes/no response. For these questions, the screen will show the question and 2 response buttons marked ‘Yes’ and ‘No’. Your task is to select the appropriate button. There are also some open-ended questions.		
When these questions are presented, you should think of an answer before clicking the ‘Next’ button. The answer could be a best guess if you cannot remember. You should consider how effortful the answer was to retrieve. Research has shown that memories that are effortful to retrieve are more likely to be wrong. If an answer was effortful to retrieve, it is probably wrong.	When you’re presented with these questions, you should review what is being asked first. After reviewing the question, you should try to retrieve an answer. More than one option could come to mind. This is perfectly normal. The answer could be a best guess if you cannot remember. Once you have retrieved your answer/s, you must make a decision about how likely it is that each possible answer is correct. You should also consider how effortful each answer was to retrieve. Research has shown that memories that are effortful to retrieve are more likely to be wrong. If an answer was effortful to retrieve, it is probably wrong. When evaluating	When these questions are presented, you should think of an answer before clicking the ‘Next’ button. The answer could be a best guess if you cannot remember.

Training	Training + retrieval	Control
	certainty and effortfulness, consider each possible answer separately. Decide which answer is most likely to be correct. This is your best answer. You may not necessarily believe that your best answer is correct and it could be a guess if you were unable to retrieve an answer. The important thing is that it is the best answer you can think of.	
You will be asked to provide two ratings for your answer. First, you will rate how effortful your memory was to retrieve on a 5-point scale ranging from 'Not at All' to 'Very Much'. Your task is to select the option that best describes your memory. Second, you will rate how confident you are that the answer is correct.		Next, you will rate how confident you are that the answer is correct.
You will be presented with a confidence scale ranging from 0-100% with 10% intervals (e.g., 0%, 10%, 20% etc.). Select the percentage that corresponds to your level of confidence. Finally, you will be asked whether you want to provide your answer. There will be two response buttons on the screen, 'Answer' and 'Don't Know'. To provide your answer in your statement to police, select the 'Answer' button. To withhold your answer from your police statement, select the 'Don't Know' button.		
When deciding whether to provide your answer or respond 'Don't Know', keep in	When deciding whether to provide your answer or respond 'Don't Know', keep in	Remember, it is absolutely essential that the information you provide is correct. Do not

Training	Training + retrieval	Control
mind that you will remember more incorrect information than you think. Consider how effortful the answer was to retrieve and remember that memories that are effortful to retrieve are usually wrong.	mind that you will remember more incorrect information than you think. Cast your mind back to when you were selecting your best answer. Consider how certain you are that the answer is correct. Consider how effortful the answer was to retrieve and remember that memories that are effortful to retrieve are usually wrong.	provide any information that could be wrong. You should only provide correct information.
When you decide to provide your answer, you will be presented with another screen containing a text box in which you can type your answer. Select 'Done' when you have finished typing your answer. When you select the 'Don't Know' button, you will proceed to the next question. You will now be given an open-ended practice question.		
PRACTICE FILTER QUESTION: Were there any customers in the bank at the time of the robbery?		
PRACTICE CLOSED QUESTION: How many customers were in the bank at the time of the robbery?		
You have now completed the practice question. Please press 'Next' to continue. Because there were two robbers, they will be referred to by their location during the robbery. One will be referred to as 'the robber by the counter'. This is the robber who took the money from the cashier. The other robber will be referred to as 'the robber off to the side'. There will be several questions about the top/s possibly worn by each robber. You will be asked if each robber was wearing a jacket, a shirt, and/or a jumper. Please ensure that you respond in the appropriate question. For example, if you remember that one of the robbers was wearing a jumper, but not a jacket, respond 'No' to the question that asks		

Training	Training + retrieval	Control
if they were wearing a jacket, and 'Yes' to the question that asks if they were wearing a jumper. If you have any questions, please ask the experimenter now.		
Please keep in mind that you will remember more incorrect information than you think. You must consider how effortful the answer was to retrieve. Memories that are effortful to retrieve are usually wrong.	Please keep in mind that you will remember more incorrect information than you think. You must always review the question being asked and retrieve all possible answers. You must consider your level of certainty, and how effortful the answer was to retrieve. Memories that are effortful to retrieve are usually wrong.	Do not provide any information that could be wrong. You should only provide correct information.
RECALL TASK (see Appendix A for list of filter and closed questions)		
You have finished the interview questions. Now we would like you to think back to the questions where you responded 'Don't Know'. Responding 'Don't Know' when you are unsure about your answer is good. However, we would like to know what your best guess was for the questions where you responded 'Don't Know'. You are going to be presented with these questions again. Please, think back to the answer you thought of when you were first presented with these questions. Do not try to come up with new answers. On each screen, there will be a question, a text box, and a button marked 'Done'. Please type the answer you were considering when you were first presented with the question and click 'Done'.		

APPENDIX F: CODING GUIDE FOR EXTERNALISED FREE-RECALL ANSWERS IN EXPERIMENT 5

Throughout the coding guide CR is used to describe the robber by the counter during the robbery and SR is used to describe the robber that was off to the side during the robbery.

Entry order x1: CR entered the bank first

Sex x2 (1 for each robber): CR was male, SR was female. They may not mention sex at all, or they may only mention it for one robber. They can state male/female or could refer to he and she. Interpret guy as male.

Age x2 (1 code per robber): Young adults, 20s-30s.

Ethnicity x2 (1 code per robber): Both were Caucasian/white.

Eye colour x2 (1 code per robber): Brown/dark for CR. Any answer for SR is wrong as this is not determinable from the video.

Hair colour x2 (1 for each robber): Any answer is wrong as no hair was seen.

Lips x1: CR had normal lips, if the say large, code as incorrect.

Face/head shape x2 (1 code per robber): CR had long/thin/slender face. No close up of SR so too difficult to judge, any answer incorrect.

Eyebrows: CR had dark (x1), bushy (x1) eyebrows. Did not see SR eyebrows so any answer is incorrect

Build x2 (1 code per robber): CR was slim/thin/average/medium. SR was bigger/chubby/fat/stocky/overweight/beer gut/wider (not average but if they say stocky/average, code as correct)

Height x2 (1 code per robber): SR was tall/fairly tall/taller. CR was short/shorter (not average). Exact height is not required but heights should be considered relative to each other if they give both (i.e., CR height should be larger than SR height). SR shouldn't be more than around 5.2 and SR should be greater but no more than 6ft (but if the relative difference is correct, code both as correct).

CR jacket: 1 code for colour (dark, dark grey, grey, charcoal, pin stripes), 1 code for type (suit, formal/sport/business jacket, blazer, button-up, just jacket), 1 code for fit (baggy), 2 codes for buttons (1 for each cuff, there were 3 button on both). If they just say suit, give 1 correct mark for jacket type and another for trouser type. If they specify the colour of the suit, give 1 mark for jacket colour and 1 mark for trouser colour. If they just say clothes were dark coloured or grey, give correct for jacket colour and trouser colour and incorrect for shirt colour (all incorrect if they say black clothing). If they say most clothes were dark or grey, give correct for jacket colour and trouser colour and no ode for shirt.

SR jumper: 1 code for colour (yellow), 1 for type (hoodie pocket in front, sports type, US flag, jumper, pullover, windcheater, sweatshirt, sweater), 1 for logo (capital/block letters, writing, USA), 1 for logo colour (dark, black), 1 for jumper sleeves (rolled/pushed up or not rolled/pushed up is correct as they were rolled up in some parts of the video and not others). Do not code if they just say top.

CR shirt: 1 code for colour (only white is coded as correct, any other colour including a light/pale colour is incorrect), 1 for type (may just say shirt but could also describe the shirt in some way such as button-up, collared, business/suit shirt).

SR shirt: no visible shirt, any answer coded as incorrect.

CR trousers: 1 code for colour (dark, dark grey, grey, charcoal, pin stripes), 1 code for type (suit pants, slacks, formal, business, formal). If they just say suit, give 1 correct mark for

jacket type and another for trouser type. If they specify the colour of the suit, give 1 mark for jacket colour and 1 mark for trouser colour.

SR trousers: 1 code for colour (dark, dark blue, blue, navy), 1 code for type (jeans, baggy, loose, long)

CR tie: 1 code for saying CR wore a tie, 1 code for tie colour (red, patterned/striped, some red, maroon, dark/deep red, some grey and white, dark)

Shoe colour x2 (1 code for each robber): Dark, black

Shoe type x2 (1 code for each robber): Closed shoes. Code anything more specific than closed shoes as incorrect. Do not give correct for just saying they wore shoes

Disguise colour x2 (1 code for each robber): Dark, black

Disguise type x2 (1 code for each robber): Balaclava, mask, ski mask, face hood/mask/cover, anything that generally describes a face mask (beanie over face with hols for eyes is ok)

Weapon x2 (1 code for each robber): Neither robber actually carried a visible weapon but made it appear as though they had weapons (probably guns). If they don't say anything about weapons, don't code anything. Any indication that they did not actually see weapons (e.g., CR hid gun in jacket pocket, SR hid gun in large bag, 'gun' in quotations) is coded as correct. If they just say there was a gun or weapon but do not mention anything about it being concealed/not visible, code as incorrect.

CR bag: CR did not have a bag. If they say he had a bag, they get an incorrect mark for bag, if they also describe the type and/or colour, they get an incorrect mark for type and/or colour, if the specifically say he did not have a bag, mark bag as correct.

SR bag: 1 code for saying there was a bag, 1 code for colour *(*black, dark, yellow or green writing), 1 code for type *(*duffel/sports/gym bag, large/long/rectangular bag, Asics), 1 code for logo (Asics). If Asics is the way they describe the bag, code correct in type and nothing for logo. If they say Asics in addition to another descriptor (e.g., sports bag) then also code correct for logo

Socks: Not visible so any answers are incorrect

Gloves: Neither had gloves so any mention of gloves is incorrect. If they explicitly say they did not wear gloves, code as correct.

Glasses: Neither had sunglasses on so any mention of glasses (regular or sun) is incorrect. If they explicitly say they did not wear glasses, code as correct.

Tattoos: Neither had visible tattoos so any indication of a tattoo is incorrect. If they explicitly state that they didn't have visible tattoos, code a correc.

Money (1 code): CR had money, it was never handed off to SR, any indication that SR had the money is incorrect. If they say that the robbers didn't get money, code as incorrect. If they don't explicitly specify who the money was handed to (i.e., cahier handed money to robber at counter or robber at counter had money), do not code.

Money location (1 code): The money was tucked away into the right side of CR's jacket. If they say this give correct mark. It was not placed in a bag, so if they say this, code as incorrect. Don't code anything if they don't specify where the money was put

Money type: Rolls of $50 notes, do not code for exact amount.

No alarm (1 code)

Exit order (1 code): SR exited bank first, followed by CR

CR escape: 1 code for each of the following (still wearing the balaclava, looked up at ceiling when leaving, stuck tongue out at camera, touched the door, ran from the bank, ran down the stairs, did not pause at bottom of stairs, headed right/east, took a bus, did not escape with bag, did not knock anyone over)

Bus: 1 code for each of the following: zone (any zone is incorrect because we didn't see the zone number. However if they say the first bus stop, this is correct), age (old), type (Adelaide metro), colour (white), number (216, started with 2 and/or had 3 digits), bus route 1 (Gepps Cross), bus route 2 (City), bus route 3 (Goodwood Road), passengers (it was not possible to tell if there were other people on the bus so code anything as incorrect), CR did not validate ticket, the bus was near trees, there was no bike on the bus

SR escape: 1 code for each of the following (still wearing the balaclava, did not look up at ceiling when leaving, did not stick tongue out at camera, touched the door, ran from the bank, ran down the stairs, did not pause at bottom of stairs, headed left/west, ran rather than taking the bus, escaped with bag, did not look back, did not knock anyone over, did not run with a limp). 1 correct if they say SR ran towards the gym/buildings, Flinders gym, Alan Mitchell sports centre/hall, lower car park (wrong if they say Sturt gym).

Direction: 1 code if the say the split up, separated or went in opposite direction and/or don't specify who went left/right (if they specify direction for one and also say split up or opposite direction, code one for each rather than the direction code)

Talking: SR did not talk (1 code), people in the bank didn't speak (1 code), there was no demand for money or request to stop and freeze (1 code)

If they say that hurry up (1 code) or sorry (1 code) were said by SR, mark as incorrect, otherwise mark as correct. 1 code if they say CR said something but do not specify what (unless they have already said that he said hurry up and sorry, then code as incorrect)

Teller who handed over money: 1 code for each of the following: female, Caucasian, age (middle aged, mid-late 30s, 40s), hair colour (brown, red/brown, dark hair), hair style (up but short), height (short), gold ring with red stone, right hand (4 codes, 1 for each detail), gold bracelet on right hand (3 codes, 1 for each detail), was not wearing glasses, short colour (white), shirt type (collar, button-up, elbow length sleeves), wore a scarf, scarf colour (patterned, grey, blue, black, navy)

Other teller at counter: 1 code for each of the following: female, Caucasian, age (middle aged, mid-late 30s, 40s), hair length (short), hair colour (dark, black hair), shirt colour (white), shirt type (collar, button up, elbow length sleeves)

Teller at back: 1 code for each of the following: male, Caucasian, age (30s), beard, hair colour (dark/black hair), hair length (short), suit, suit colour (dark/black), shirt colour (white), shirt type (business shirt, collared, button-up, can just say shirt), tie, tie colour (grey, blue).

Customer: 1 code for each of the following: female, Caucasian, age (young, 20s), hair colour (dark, brown, red/brown hair, auburn), hair was up, top colour (white), top type (sleeveless, collared, could just say shirt), skirt colour (light, pale, pink, baby pink), skirt type (knee-length or just skirt), shoe colour (dark, brown, black), shoe type (slip on, open back), covered her face

Number of tellers: 3, if they describe all three, give correct for this, they may also just say there were 3, if they say there were less than three, code as incorrect, if they give a range that contains the true value (e.g., 1-4), give correct

Number of customers: 1

Total number of witnesses in the bank: 4 (only code for this if they have not specified between workers and customer)

Witnesses present: if they just say there were other witnesses in the bank and don't specify the amount code correct

Witness on stairs: 1 code for each of the following: male, Asian, dark/black hair, sunglasses (can just say glasses)

Bank: ANZ

Location: Flinders University

Time frame: if the give a range than contains the correct answer (1 minute), give correct

Time of day: day time, mark correct if they say it was sunny

Injuries: none

Threats: none

Scream: no one screamed (code the separate from whether witnesses talked)

Register type: wood

Street: street was quiet rather than crowded

The bus was near an electric box

Electric box was green