ABSTRACT

Weakening the "Illusion of Memory Knowledge":
A Potential Method for Improving Jurors' Evaluation of Eyewitness Evidence

Courtney A. Kurinec, Ph.D.

Mentor: Charles A. Weaver III, Ph.D.

Many efforts have been made to educate jurors about factors that influence the reliability of eyewitness memory. However, most of them fail to improve jurors' sensitivity to the quality of a particular memory and instead only induce skepticism of all eyewitness evidence. One potential method for improving a prospective juror's understanding of memory may be to ask them to explain a concept before providing expert information. By forcing jurors to acknowledge the limitations of their knowledge about memory, they may be more attentive to new, more accurate information. Over three experiments, I assessed whether explanations lead to a reassessment of memory knowledge (Experiment 1), explored whether reassessing one's understanding leads to improved metacognition and learning (Experiment 2), and investigated the applicability of this task to the courtroom (Experiment 3). The findings from this project provide initial evidence that engaging in explanations may be a low-cost method for improving jurors' evaluation of eyewitness evidence.

Weakening the "Illusion of Memory Knowledge":
A Potential Method for Improving Jurors' Evaluation of Eyewitness Evidence

by

Courtney A. Kurinec, B.A., M.A.

A Dissertation

Approved by the Department of Psychology and Neuroscience

_____

Charles A. Weaver III, Ph.D., Chairperson

Submitted to the Graduate Faculty of
Baylor University in Partial Fulfillment of the
Requirements for the Degree
of

Doctor of Philosophy

Approved by the Dissertation Committee

_____

Charles A. Weaver III, Ph.D., Chairperson

_____

Michael K. Scullin, Ph.D.

_____

Wade C. Rowatt, Ph.D.

_____

Jo-Ann C. Tsang, Ph.D.

_____

Tracey N. Sulak, Ph.D.

Accepted by the Graduate School

August 2019

_____

J. Larry Lyon, Ph.D., Dean

Page bearing signatures is kept on file in the Graduate School.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I am incredibly thankful for the support I have received from friends, family, and colleagues over the course of my graduate school career. First and foremost, I would like to thank my mentor, Dr. Charles Weaver, for his advice, encouragement, and humor (often at my expense). I am also thankful for my committee members for their guidance and support on this and other projects: Dr. Michael Scullin, Dr. Wade Rowatt, Dr. Jo-Ann Tsang, and Dr. Tracey Sulak. I would be remiss if I did not thank the faculty, staff, and my fellow graduate students in the Department of Psychology and Neuroscience and in the Academy for Teaching and Learning for teaching me, supporting me, and most of all bearing with me. I would especially like to thank all the members of my laboratory, both past and present, for their hard work and assistance. Finally, I am grateful for my family and friends, who saw me through my graduate education. Thank you all for being a part of my life.

DEDICATION

To everyone and everything that kept me going:

Family, friends, cats, coffee, plants, and sunshine

CHAPTER ONE

Background and Significance

Despite continuing advancements in the scientific understanding of memory, members of the general public remain largely under-informed of how memory works. Several surveys over the last decade indicate that members of the public have an inflated sense of the reliability of memory. A 2009 nationwide telephone survey on general memory beliefs revealed that 82.7% of respondents incorrectly believed that amnesia resulted in a person failing to remember his/her own name, 63.0% incorrectly believed that memory works like a video camera, and 47.6% incorrectly believed that memories are permanent and unchanging (Simons & Chabris, 2011). These results were later replicated in a 2011 online survey (Simons & Chabris, 2012), suggesting that a scientific understanding of memory has not yet reached the layperson.

These findings are of particular concern when considering the role memory can play in the courtroom, where members of public serve as members of the jury. If jurors believe memories to be permanent and unchanging, why would they doubt the veracity of an eyewitness's testimony, particularly one that has no apparent ulterior motive to misrepresent the facts? This indiscriminate confidence in the reliability of memory can cause jurors to accept demonstrably false identifications and recounts of events, ultimately resulting in judgments against innocent persons. It is unsurprising, then, that eyewitness misidentification remains a leading cause in false convictions. In nearly 70% of cases later exonerated by DNA evidence, eyewitness misidentification contributed to

the decision to convict (Innocence Project, 2019a). The human and financial costs of these false convictions are great – according to the Innocence Project (2019b), the 364 people exonerated due to DNA evidence since 1989 spent on average 14 years behind bars before their exonerations. Convicting the wrong individual clearly harms a specific family and community, but also affects those completely removed from the case. For example, false convictions have cost Texas taxpayers $93.6 million over the past 25 years (Silver & Carbonell, 2016); the California state government has paid out $282 million over the course of 24 years (Jackman, 2016). Concerningly, false convictions also mean that the true perpetrator remains at large.

Despite the serious consequences associated with doing so, most people seem unaware of the risk of blindly trusting eyewitness testimony. Surveys on more specific eyewitness memory knowledge indicate that prospective jurors' beliefs about memory mostly deviate from those of memory experts (see Benton, Ross, Bradshaw, Thomas, & Bradshaw, 2006). A 2004 survey of potential D.C. jurors found more than 66% of respondents believed they would never forget a face, and 46% incorrectly believed that "the witness on the stand is effectively narrating a video recording of events that she can see in her 'mind's eye' for jurors" (Schmechel, O'Toole, Easterly, & Loftus, 2006, p. 196). Other incorrectly endorsed items included the belief that witnesses would be more reliable in instances where a weapon was present (37%), that violence would increase memory reliability (39%), and that a witness' level of confidence was a reliable marker of reliability (39%). Clearly, the gap between current research and the average person's understanding of how memory works must be addressed, especially for prospective jurors.

Many efforts have been put forth in an effort to educate jurors on factors that influence the reliability of memory. However, many of them fail to improve jurors' sensitivity to the quality of a particular eyewitness's memory; instead, most methods serve only to induce skepticism about the reliability of all memories (e.g. Papailiou, Yokum, & Robertson, 2015). The current research investigated the viability of a potential method for improving a layperson's assessment of their own understanding of memory and perhaps subsequent learning on the subject. Before describing the three experiments that examined this new method on improving calibration, learning, and its application in the courtroom, however, a review the literature on eyewitness memory and overconfidence is in order.

Eyewitness Memory

Evaluating the reliability of an individual person's memory is not a straightforward process; numerous factors influence the dependability of memory. When discussing eyewitness memory in particular, these factors are generally divided into two broad categories: system and estimator variables. System variables refer to those factors that potentially can be controlled by the legal system and include factors such as the style of interview process and lineup procedures (Wells, Memon, & Penrod, 2006). Estimator variables, on the other hand, are those out of the control of the legal system, such as the lighting conditions when an eyewitness is viewing an event or whether the perpetrator wore a hat. Although not every variable known to influence the quality of eyewitness memory is relevant to the present research, I will provide a brief overview of several variables to demonstrate the extent to which eyewitness memory can be altered.

System Variables

Despite recent efforts to introduce standardized witness interrogation and suspect lineup procedures intended to limit the influence of system variables on eyewitness identifications (see National Institute of Justice, 2003; National Research Council, 2014), these recommendations have not been uniformly adopted across the United States (Innocence Project, 2019c). Further, even within offices that have adopted the measures, there are no guarantees that these measures are being employed as recommended. Given the difficulty in regulating every interaction between a law enforcement officer and a witness, choices made by agents of the criminal justice system likely continue to affect the reliability of eyewitness memory. These choices, like who to include in a lineup, how to give instructions, and the information given either explicitly or implicitly to eyewitnesses, can all influence not only the identifications an eyewitness makes, but also the confidence an eyewitness has in those identifications.

Mugshot bias. When a witness is asked to make an identification, they may be shown a mugbook containing photographs of suspects from previous crimes before any individuals are brought in for lineup or showup procedures. However, viewing these mugshots may lead to what is known as mugshot bias, or a preference to choose those suspects who have previously been seen. Memon, Hope, Bartlett, and Bull (2002) found that witnesses who viewed and selected a suspect from a target-absent mugbook were more likely to choose a suspect who appeared in both the mugbook and a later target-absent lineup, regardless of who they chose from the mugbook originally. Those witnesses who saw the mugbook but did not identify a suspect did not show a preference for the repeated suspect. Although some researchers have not found evidence for a

mugshot bias effect (see Lindsay, Nosworthy, Martin, & Martynuck, 1994) a meta-analysis by Deffenbacher, Bornstein, and Penrod (2006) found that exposure to mugshots between encoding of the perpetrator and identification can decrease the proportion of correct identifications and increase the proportion of false identifications. Although it seems logical to show witnesses the same suspects multiple times, especially if they are relevant to the crime in question, showing these images risks contaminating an eyewitness's memory for the perpetrator.

Lineup instructions. Another way law enforcement officials can influence eyewitness identification is through the use of biased lineup instructions. When witnesses are asked to make an identification, they may be unaware that the perpetrator is not necessarily in the lineup. Witnesses often want to aid law enforcement, and demand characteristics may lead them to identify a suspect from a given lineup even if they are unsure because they trust that officers have found the suspect. If witnesses receive biased lineup instructions – in other words, they are not informed during identification procedures that the suspect may not be present in the lineup, and that they have the option to refrain from identifying a suspect – they can feel pressured into choosing a suspect. In a meta-analysis of 18 articles that manipulated biased versus unbiased lineup instructions, Steblay (1997) found that biased instructions led participants to make a choice from the lineup more frequently. This pressure to choose is of particular importance when considering research that suggests that unbiased lineup instructions lead to correct identifications more often than biased lineup instructions (see Leippe, Eisenstadt, & Rauch, 2009), particularly when the target suspect is absent from the lineup.

Post-event information. Law enforcement officials must also be cautious with the information they provide to the witness after the event. Post-event information, which is not necessarily misleading or intentional, can alter a witness's confidence in their identifications or even their memory for the event itself. Post-event information included in questions about the event can lead witnesses to endorse false information, regardless of whether the false details are peripheral (Bowers & Bekerian, 1984; Loftus, Miller, & Burns, 1978) or central to the event (Köhnken & Brockmann, 1987; Loftus, 1975; Loftus & Palmer, 1974). Even something as seemingly innocuous as information about who co-witnesses have identified or details from the co-witnesses themselves can alter a witness's level of confidence in their identifications (Luus & Wells, 1994). Co-witness information can even alter an eyewitness's memory for the event, particularly when the co-witness endorses incorrect information (Shaw, Garven, & Wood, 1997).

Confidence malleability. Finally, even if law enforcement are cautious in how they handle identifications and avoid providing biasing information, they must also be careful when responding to witness identifications. Lineup administrators and other law enforcement officials can, explicitly or implicitly, influence an eyewitness's confidence in their identification. When witnesses receive indications that their identification matches the suspicions of the administrator, through either verbal or non-verbal confirmations, they become more confident in their choice (Garrioch & Brimacombe, 2001; Wells & Bradfield, 1998). A meta-analysis by Douglass and Steblay (2006) found that participants who received confirmatory feedback for their suspect identifications had greater confidence in their identifications and were more likely to report better viewing conditions than those who received no feedback. In addition to being more confident,

these witnesses were also more willing to testify. Given that no actual improvement in their memory for the perpetrator occurred, such testimony may give jurors a false impression that the identification is more reliable than it truly is.

Estimator Variables

Factors outside the purview of law enforcement pose an arguably greater threat to the validity of eyewitness memory. Estimator variables cover a variety of factors, from perpetrator characteristics (e.g. race, wearing a disguise, carrying a weapon) to eyewitness characteristics (e.g. stressed, intoxicated, age, familiarity with the perpetrator) to characteristics in the environment surrounding the event (e.g. lighting, interference). Although these factors cannot be controlled by law enforcement, the "understanding of the importance of any given system variable is, at least at the extreme, dependent on levels of the estimator variables" (Wells, Memon, & Penrod, 2006, p. 51). Thus, understanding how estimator variables can influence the quality of an eyewitness's memory for a perpetrator can help determine the impact system variables have, if any, on the accuracy of that identification.

Cross-race bias. Whether the race of the eyewitness and the perpetrator match can affect the accuracy of an eyewitness's identification. A meta-analysis of 39 articles comparing own- and other-race bias found a significant effect of own-race bias – people were nearly 1.5 times more likely to correctly identify a target face of their own race (Meissner & Brigham, 2001). A cross-race effect also emerged, such that people were nearly twice as likely to misidentify a novel face as the target if the face belonged to a member of another race. Together, these findings suggest that people are better at

discriminating between faces of their own race than other races, an effect which may increase the risk of false identifications when a witness is identifying someone from another race or ethnic group.

Weapon focus. The presence of a weapon during the event can also weaken the reliability of an eyewitness's memory for an event. According to the Bureau of Justice Statistics, around 22% of all violent crimes in the United States in 2009 involved a weapon (Truman & Rand, 2010). Over two experiments, Loftus, Loftus, & Messo (1987) found that participants who viewed photos of a crime that included a weapon were more likely to fixate on the weapon and less likely to identify the perpetrator compared to those who saw no weapon. Because individuals focus their attention on a weapon, either due to the threat it poses or due to its unexpectedness, they impair their ability to fully encode details about the person holding that weapon, undermining the accuracy of their memory and their identifications.

Alcohol intoxication. In addition to suspect-specific factors like race and attire, the state of the eyewitness at the time of the event can also impact the reliability of memory. A witness's level of blood alcohol concentration (BAC) can impair their ability to correctly encode information about the event. As BAC increases, people are less likely to correctly identify target individuals – particularly in target-absent lineups – and may also suffer impairments to their memory for details of what they themselves did (Dysart, Lindsay, MacDonald, & Wicke, 2002; Read, Yuille, & Tollestrup, 1992). However, despite appearances, an eyewitness who has been drinking should not be dismissed entirely. When under conditions of high arousal, people who consumed enough alcohol to

have BAC readings around .08 (the legal limit in the United States; Insurance Institute for Highway Safety, 2017) were just as likely to identify an individual central to the crime as those who had a placebo drink (Read, Yuille, & Tollestrup, 1992).

Exposure duration. Other factors which affect the quality of an eyewitness's memory are entirely dependent on the context of the event itself. For example, eyewitnesses who are exposed to a crime or a perpetrator for longer periods of time are more likely to recall those details than eyewitnesses with only brief exposure. Memon, Hope, and Bull (2003) asked participants to view a short video of a robbery in which the robber was visible for either 12 seconds or 45 seconds. After over a half-hour of filler tasks, participants were asked to identify the robber from a simultaneous lineup. Those participants who saw the target for a longer amount of time were more accurate and more confident in their lineup decisions. Further, those in the short exposure condition were more likely to make an incorrect identification. These findings suggest that exposure duration is positively related to positive identifications.

Unconscious transference. Even if all the necessary conditions for a reliable identification are in place, witnesses may still misidentify a perpetrator if they failed to encode information about the perpetrator's appearance accurately. People can and often do fail to notice changes in a person's appearance. In fact, Simons and Levin (1998) demonstrated that people can even sometimes fail to realize that the person they were interacting with had been replaced with someone else. These errors occur because people often encode only general, categorical or stereotypical information about the people they encounter (e.g. "construction worker"). As a result, people may unconsciously transfer

the details they do recall about an individual to a similar but unrelated person. In the context of eyewitness identifications, the consequences of such a mistake may be severe. Of concern, Ross, Ceci, Dunning, and Toglia (1994) found that witnesses may still misidentify an innocent individual as the perpetrator even after being made aware they were viewing a crime and told to attend to the perpetrator. Participants who viewed an innocent bystander at the beginning of a crime video were more likely to identify him as the perpetrator than those who did not view the bystander. Witnesses who saw the bystander conflated their memory of the two men, concluding that the bystander was familiar because he had committed the crime. Thus, even under relatively ideal circumstances (e.g. aware that a crime was taking place), witnesses can still misidentify the perpetrator if they are unable to determine why an individual appears familiar.

Confidence

No review of eyewitness memory would be complete without mentioning confidence. Although an eyewitness's stated confidence doesn't necessarily impact the accuracy of a given eyewitness's identification, the influence on jurors cannot be understated. Therefore, I will provide an overview of the role confidence plays with regards to eyewitness testimony both at identification and at trial.

Previous experimental research found the relationship between confidence and accuracy to be either small and positive (Sporer, Penrod, Read, & Cutler, 1995) or not statistically significant at all, especially for less than ideal viewing conditions (Deffenbacher, 1980). However, a growing body of literature asserts that stated confidence can indeed be a good indicator of accuracy, provided that conditions at the time of testing are unbiased, the duration between the event and testing is relatively short,

and witnesses give their confidence ratings shortly after they make a positive identification (Brewer & Wells, 2006; Palmer, Brewer, Weber, & Nagesh, 2013; Wixted & Wells, 2017). Though the confidence a witness presents at trial may not always be indicative of a witness's accuracy (e.g. Bothwell, Deffenbacher, & Brigham, 1987), indications of extreme confidence at the time of identification are likely to be informative.

As previously discussed, confidence is malleable and can be influenced by post-event feedback or other memory distortions. Yet in the courtroom, jurors are overwhelmingly insensitive to this distinction. Jurors are more likely to render their verdicts in accordance with a highly confident eyewitness than one who hesitates or displays nervous behaviors (Bothwell & Jalil, 1992; Brewer & Burke, 2002). Outside of the courtroom, prospective jurors heartily endorse the confidence-accuracy link. In a survey of registered Florida voters, 56% incorrectly believed confidence reliably predicted accuracy (Brigham & Bothwell, 1983), and time has not advanced the modern juror's knowledge on the subject. A 2004 telephone survey of potential jurors in the District of Columbia found that 40% of respondents believed that confidence was "an excellent indicator of that eyewitness' reliability" (Schmechel, O'Toole, Easterly, & Loftus, 2006, p. 199). Thus, although the scientific community is aware of the nuance regarding confidence's predictive ability and has made recommendations accordingly (see National Research Council, 2014), the average juror is unaware of this distinction.

If all these factors can each affect the quality (or apparent quality) of an eyewitness's identification, how can jurors possibly be expected to learn, understand, and

apply relevant information in the courtroom, when their time and mental resources are already maximized?

Attempts to Inform Jurors about Eyewitness Memory

The scientific community has attempted to address this disconnect between juror and expert knowledge through the use of expert eyewitness testimony admitted at trial. For the most part, the decision to admit expert testimony is generally based on two standards: *Frye* and *Daubert*. The *Frye* standard, established in *Frye v. United States* (1923), permits expert testimony on topics unlikely to be common knowledge to the jury, provided that the testimony conveys the consensus opinion of the respective scientific community. Although many states continue to rely on the *Frye* standard for admitting expert testimony (Jurilytics, 2017), federal courts and some states have instead adopted the *Daubert* (1993) or similar standard. Under the *Daubert* standard, it is up to the trial judge to determine if the expert's testimony is based on valid, reliable data and methods. This broad directive allows scientific evidence to be presented in context but places a large amount of discretion on the shoulders of the judge, who is as likely to be unfamiliar with the subject matter as the jury themselves (Walsh, 1998). Although many judges still believe that expert testimony on memory is unnecessary, as this information is "not 'beyond the ken' of the average juror" (Schmechel, O'Toole, Easterly, & Loftus, 2006, p. 177), a number of rulings in recent years have come out in favor of permitting expert testimony on eyewitness memory (e.g. *Commonwealth v. Walker*, 2014). These courts refer to the extensive aforementioned research on eyewitness memory, noting that "eyewitness identifications are potentially unreliable in a variety of ways unknown to the average juror" (*State v. Guilbert*, 2012, p. 739).

When permitted, expert testimony on eyewitness memory clearly can be beneficial for the jury. Experts have a deep knowledge of the current literature and research consensus, and they can focus their testimony to focus on those factors most relevant to the case (National Research Council, 2014). Further, experts can be cross-examined, which allows the testimony to be challenged and gives the jury additional context with which to evaluate the information. However, the use of expert testimony is not without its own problems. Use of expert witnesses can create issues due to the increased time for the expert's testimony and high cost of hiring an expert for many defendants (National Research Council, 2014; Simmonsen, 2011). According to the Bureau of Justice (Harlow, 2000), in the 1990s around 66% of felony defendants in Federal courts and 82% in State courts were indigent, or unable to afford their own representation and thereby represented by court-appointed counsel. When courts do not make policies to provide funding toward the use of expert witnesses, defendants in criminal cases may be unable to afford to retain these experts.

Additionally, the use of an expert witness at trial may overwhelm jurors with information, and there may be conflicting expert accounts, complicating jurors' understanding of the technical information presented (Simmonsen, 2011). Moreover, an expert can sometimes be perceived as a "hired gun" and thus perceived as less trustworthy, especially when the expert receives higher pay, when it's known that they frequently serve as an expert, and when the testimony is complex (Cooper & Neuhaus, 2000). As a result, many in the scientific and legal community advocate instead for the use of jury instructions to inform jurors about eyewitness memory. Pattern jury instructions are already commonly used in trial to educate jurors about the law for certain

crimes or legal situations (see Eleventh Circuit, 2016; Fifth Circuit, 2015); therefore, the addition of memory rebuttal information should not drastically alter court proceedings.

Although jury instructions were presumed effective at educating jurors about memory, there have been concerns about how well jurors are able to employ what they learn from these instructions to assessments about the quality of an eyewitness's identification. Recent research suggests that jury instructions in their suggested form do not improve jurors' ability to appropriately apply their knowledge; instead, the use of instructions increases jurors' skepticism indiscriminately (Berman, 2015; Papailiou, Yokum, & Robertson, 2015) or have no effect on jurors' sensitivity to eyewitness conditions (Jones, Bergold, Dillon, & Penrod, 2017). The jurors themselves are remarkably unaware that they are not using what they have learned – in one study, most jurors indicated that they understood the instructions (Papailiou, Yokum, & Robertson, 2015). This overconfidence in understanding highlights a larger problem facing memory experts in the courtroom: the fact that people are often not well calibrated in their own knowledge.

Overconfidence

> *[T]o the extent that people do not have perfect memory, and the imperfections include both gaps and mistaken information, or even information that is close to correct but not exactly correct, their metamemory judgments are bound to be both inaccurate and systematically biased.* (Metcalfe, 1998, p. 106)

What a person knows about their own memory and cognition is known as metacognition (Flavell, 1979). Metacognition is critical for monitoring and controlling one's performance on a given task, and therefore requires an understanding of not only the knowledge necessary to complete the task, but also the skills to assess progress and

adjust accordingly (Nelson & Narens, 1990). Early conceptualizations of the *meta* or abstracted aspects of memory assumed that individuals were accurate in their assessments of their own memory ability; however, as previously discussed, memory is susceptible to suggestibility and does not necessarily – and is in fact unlikely to be – representative of an objective truth (O'Sullivan & Howe, 1995). As a result, current theoretical models acknowledge that individuals can be flawed in their assessments.

One type of these flawed assessments is overconfidence. Overconfidence occurs when people's judgments about their performance exceed their actual performance. However, some argue that what is generally perceived as overconfidence is in fact three separate phenomena: overestimation of one's overall performance, overplacement of one's own rank compared to others', and overprecision with the level of confidence in the accuracy of one's answer (Moore & Healy, 2008). Although the nature of these three effects are discrete, the literature commonly refers to all judgments of this nature as overconfidence. As these distinctions are not relevant for the current research, the umbrella term *overconfidence* will be used to broadly refer to all three of these phenomena.

Arguably one of the most worrisome findings regarding overconfidence is that under-skilled individuals are less able than more skilled individuals to recognize their own lack of ability. Over four studies, Kruger and Dunning (1999) found that those who received scores in the lowest quartile generally rated their performance as above average. This lack of calibration between what they actually knew and what they believed they knew occurred across multiple domains, including humor, logical reasoning, and grammar. Further, this overconfidence effect persisted in the face of objective

information on their peers' performance. Participants in the bottom quartile did not change their original estimates of their own performance, even after viewing the work of their peers, suggesting that underskilled individuals may be unable to effectively use social comparison to make more accurate judgments of their performance.

Despite appearances, overconfidence does not necessarily increase linearly with performance. Though Kruger and Dunning (1999) found that the worst performers were the most overconfident, they also noted that those at the top tended to underestimate their performance. In a series of five experiments, Lichtenstein and Fischhoff (1977) found that performance and accuracy followed a curvilinear relationship; performers in the lowest quartile were the worst calibrated, but the very top performers tended to be underconfident. In fact, those participants with around 78% correct who were the most calibrated. Therefore, although overconfidence is a problem plaguing low performers, the best performers also struggle with their metamemory judgments.

The overconfidence effects observed in laboratory experiments are not just abstract assessments – the belief in one's non-abilities can inform one's decision, no matter how erroneous. Fischhoff, Slovic, and Lichtenstein (1977) found that participants were willing to bet real money based on their erroneous confidence. Participants were given a list of questions from various domains. Each question had two possible choices, and participants were asked to indicate the correct answer, as well as provide the odds that their answer was correct. After completing the questionnaire, participants were encouraged to participate in a gambling game where the experimenter would draw a chip out of a bag for every answer that participants believed had 50:1 or greater odds of being correct. If participants were incorrect, they would pay the experimenter $1. Additionally,

if the experimenter drew a red chip out of the bag (50:1), participants would be paid $1. When given the opportunity to participate in a gambling game for actual money, participants did not want to go back and change their responses – participants remained overconfident in their incorrect answers, despite the fact that in reality they would have lost money on average.

Why are people so miscalibrated when making judgments of their own knowledge that they would risk real resources? The problem may lie in the cues they are using to assess their level of knowledge. According to the cue-utilization theory (Koriat, 1997), people rely on intrinsic, extrinsic, or mnemonic cues to inform their assessments of their own knowledge. Intrinsic cues are those cues which serve as indicators of difficulty, like the strength of semantic association between paired associates or the imageability of a single word. Extrinsic cues, on the other hand, are related to the conditions of learning or encoding processes utilized, such as the type of practice (e.g. massed or distributed) or the levels of processing required for learning. Together, these two cues in turn inform mnemonic cues, or the subjective indicators of how well something has been learned.

Although intrinsic and other experience-dependent cues are generally less predictive of future performance, people tend to discount the value of extrinsic compared to more experiential cues (Kelley & Jacoby, 1996; Koriat, 1997). Instead, people rely on cues like retrieval fluency (Kelley & Jacoby, 1996) and familiarity (Metcalfe, Schwartz, & Joaquim, 1993) to inform their judgments of how well something has been learned. It is unlikely that people are using these misleading cues in an effort to deceive themselves; rather, people have likely been taught to rely on the wrong cues to inform their assessments of their own knowledge.

However, people may not be as unaware of their own performance as previously suggested. Son and Kornell (2010) investigated whether there were any "benefits of recognizing one's own ignorance" (p. 210). Students were asked to make judgments of learning (JOL) as they studied a series of word pairs. Afterward, participants were asked to choose the pairs they would like to study more. Participants overall chose to restudy items that they rated as most difficult, indicating that participants were aware of those items on which they would likely perform the most poorly. However, only half of the participants were allowed to restudy their chosen pairs; the rest were only allowed to restudy the items they hadn't selected. Those who restudied the items they had chosen performed better than those who were prevented from doing so, suggesting that people are somewhat aware of what they don't know, at least in general.

Although the extent of overconfidence may not be as drastic as earlier research suggested, overconfidence continues to pose a problem insofar as it influences people's decision making and other behaviors. Given that jurors are particularly under-informed about memory, yet likely do not believe themselves to be so when making their assessments about an eyewitness's memory, the problem of overconfidence cannot merely be identified and reviewed; solutions to this problem must be proposed to avoid the negative consequences of jurors' over-reliance on eyewitness memory. How, then, can the blindness caused by overconfidence be overcome?

Improving Metacognitive Calibration

Addressing the problem of miscalibration has long been a goal of the research on overconfidence. In fact, even early research on the overconfidence effect examined potential solutions (e.g. Lichtenstein & Fischhoff, 1977). Although some methods with

promising results have been identified, no one solution has been discovered to improve metacognitive accuracy. I will discuss two of these methods, training and delaying judgments, below.

Training

Training intended to improve calibration has received mixed results. Lichtenstein and Fischhoff (1977) found that training on task resulted more accurate answers and better calibration overall compared to no training and that receiving training on how to accurately utilize confidence probabilities improved calibration, even after just one session (Lichtenstein & Fischhoff, 1978). Kruger and Dunning (1999) found similar results, particularly for those who were the least competent – those in the bottom quartile who received training improved to the point they were similarly accurate in their judgments as those in the top quartile who did not receive training.

However, calibration training may not necessarily generalize to other tasks, regardless of their similarity to the training task (Lichtenstein & Fischhoff, 1978). Even if training provided a consistent effect for some tasks, it is impractical, if not impossible, for individuals to receive assessment training on their everyday knowledge. As a result, it may be more efficient to simply receive feedback to serve as a form of task training; further, feedback has been found to improve learning (e.g. Pashler, Cepeda, Wixted, & Rohrer, 2005). However, these improvements may also be limited. Dunlosky and Rawson (2012) asked participants to study several terms and definitions to criterion. Participants were asked to self-score their responses; half of participants received a list of idea units to check off before self-scoring, whereas the other half received no such guidance. Although participants who received feedback in this initial study were less overconfident during the

study phase, feedback did not change the outcome for those who were overconfident in a follow-up study where all participants received the idea-units. Participants who performed the worst were the most overconfident, and as overconfidence increased, performance during the study phase and on the final test decreased. These differences in performance due to overconfidence persisted even when accounting for participants' learning ability, and there were no differences in study time or number of trials before reaching criterion. These results suggest that feedback alone may not be enough to overcome people's own miscalibrated judgments of accuracy.

Although training and feedback may be useful tools in other domains, it is unlikely that jurors would have time to "train" before trial or that they would receive feedback on their judgments about the case. Further, it is even more unlikely that an incompetent individual would be self-aware enough to identify their lack of skill in order to sufficiently seek help on it. Therefore, although training and similar methods may be beneficial in specific contexts, its utility is severely limited in the courtroom.

Delaying Judgments

Research on delayed judgments has produced far more consistent results: delaying when people make assessments about their knowledge leads to more calibrated judgments. After giving participants a series of word pairs to learn, Nelson and Dunlosky (1991) found that JOL made following a delay resulted in Goodman-Kruskall gamma coefficients (a measure of rank correlation, in this case between judgments and accuracy) of .90, a nearly perfect relationship. Comparatively, those JOL made immediately after learning had coefficients of only around .36. Although the cause of this improvement has been widely debated, research conducted in the lab has found that the effects due to delay

are not merely the result of short-term memory recency effects (as claimed by the Monitoring-Dual-Memories hypothesis; Nelson & Dunlosky, 1991), nor are they simply the result of the polarization of ratings usually seen when making judgments at delay (proposed by Dunlosky & Nelson, 1994). Kelemen and Weaver (1997) found that both distraction intended to disrupt short-term memory and delay improved calibration coefficients compared to immediate JOL, but delay produced the highest gamma coefficients. Separately, Weaver and Kelemen (1997) examined the influence of the more extreme distributions of JOL on gamma coefficients and found that, although the distribution of responses using the extreme ends of the scale did increase ratings of calibration, the improvement in calibration due to delay of judgment was greater.

Overall, delaying judgments improves calibration over and above any changes due to short-term memory monitoring or the shape of the distribution of responses. Yet, when applied to the courtroom it seems unlikely that jurors would delay their judgments about any one piece of evidence. According to the story model of decision making (Pennington & Hastie, 1992) jurors rely on narratives created over the course of the trial to help them organize and evaluate case information. Jurors then use these stories to inform their judgments. As jurors receive a large amount of information, they are unlikely to pause before making any judgments – instead, jurors likely incorporate their perceptions of credibility and culpability as they observe the proceedings, perhaps without their own awareness. Even with conscious awareness, jurors may not be able to stop themselves from making these assessments. Therefore, any attempt to address overconfidence in the courtroom must rely on a methodology that can be easily

administered, both under the constraints of the location and the jurors' own cognitive abilities.

Explanations and the Weakening of the Illusion of Explanatory Depth

Given the limits for operating in the courtroom, overtly asking people to provide explanations may be a more viable method for improving calibration. Previous work by Koriat, Lichtenstein, and Fischhoff (1980) found that participants were less overconfident and slightly more correct when they were required to provide reasons for their answers. Importantly, participants were the most calibrated when required to provide reasons against their chosen answer; providing supporting reasons alone did not differ from providing no reasoning.

The findings by Koriat et al. (1980) suggest that people tend to give less weight to information that contradicts their position and therefore fail to properly consider alternatives to their response, leading to overconfidence. Giving explanations against their own answer forced them to consider these possibilities, improving their overall calibration. However, people may be overconfident despite their purported understanding because they are reasoning using their own personal, flawed knowledge of how a certain phenomenon works as a basis for these assessments. According to Keil (2003), people sometimes confuse knowing *how* something works on a functional level with how it operates or is assembled internally. Additionally, people sometimes conflate knowing where information on a certain topic is stored or how to access it with true knowledge on the topic. Although interactive, external memory stores such as experts or the Internet are clearly beneficial to the individual – the individual has access to more sources of knowledge – these transactive memory systems can result in overconfidence (Wegner,

1987). For example, Fisher, Goddu, and Keil (2015) found that people report higher levels of knowledge on a topic when they perform Internet searches on it, even if the searches did not result in an answer. Separately, Sparrow, Liu, and Wegner (2011) found that people tend to forget the actual information when they know it can be accessed electronically later, yet are better at recalling where that specific information can be accessed. Simply knowing that an answer can be found later can result in illusions of understanding on the subject in question.

Although these illusory frameworks about the depth of knowledge appear problematic at first glance, they are beneficial and perhaps necessary for everyday life. These frameworks allow people to see coherent, causal patterns without needing to know everything at a more specific level. Further, these loose frameworks can be quickly applied to new information; for example, cognitive schema allow people to feel they understand similar situations even if they have never encountered a specific situation before. Through these frameworks, people are able to navigate the world around them without needing to know every detail about their environment.

As a result of these loose mental structures, people feel they have a good understanding of how things work, despite not having taken the time to analyze the depth of their understanding properly. This limited knowledge about how things work while lacking the realization that one's own knowledge is in fact shallow is referred to as *the illusion of explanatory depth* (IOED; Rozenblit & Keil, 2002). Weakening this illusion by asking participants to explain how devices and phenomena work causes them to reassess their original level of understanding to be more calibrated with their actual level of knowledge. Over the course of twelve studies, Rozenblit and Keil (2002) demonstrated

the existence of the IOED and the effects of weakening it. Participants were first instructed and trained on a 7-point Likert-type scale (1 = Lowest understanding, 7 = Highest understanding). In the training process, participants were shown an example of explanations from the target domain (e.g. mechanical devices) which demonstrated a shallow, intermediate, or deep understanding and their associate rank on the scale. Next, participants were shown a list of items and asked to rate their level of understanding without "pausing excessively on any item" (p. 527). After providing their initial ratings, participants were asked to write an explanation for a subset of the list of items (e.g. "how a helicopter changes from hovering to forward flight"; "how earthquakes occur") before re-rating their understanding. Finally, participants were asked to describe how each item worked in response to step-by-step questions before once again rating their understanding. When participants experienced a weakening of the IOED, for example, for devices, they re-rated their understanding as lower than their initial rating. Promisingly, these ratings are not merely lower due to changes in self-esteem or changes in how the rating scale is perceived; studies where these explanations were rated by strangers found that participants are fairly accurate when rating their post-explanation level of understanding (Alter, Oppenheimer, & Zemla, 2010).

Unlike more traditional general measures of overconfidence, the utility of ratings used with the IOED paradigm varies depending on the domain in question; that is, on what is being explained. Rozenblit and Keil (2002) found a large decrease in ratings of understanding for devices and natural phenomena, a small decrease in ratings of understanding for facts, and no change in ratings for narratives or procedures. Consequently, level of transparency of a device, e.g. the ratio of hidden to visible parts,

was the best predictor for the extent of overconfidence in participants' ratings of understanding. Therefore, any weakening of IOEDs should ideally occur in domains that are more abstract and whose inner workings less clear. Accordingly, Alter, Oppenheimer, and Zemla (2010) found that people who are primed to think more abstractly are more likely to experience a weakening of an IOED. If people are encoding only loose causal frameworks to explain the world around them, they are more likely to be overconfident in their understanding of more abstract domains like devices than more concrete ones like facts or procedures.

Although any weakening of IOEDs is limited to those domains which are more obscure with regard to their inner workings, the effect operates in fields other than those examined by Rozenblit and Keil (2002), to include psychology. Zeveney and Marsh (2016) asked participants to rate a series of questions regarding devices (e.g. "How a zipper works") or mental disorders (e.g. "How the different symptoms of depression develop") before being asked to give explanations and re-rate their perceived understanding. Participants also rated the degree to which they believed other members of society understood the subject and the perceived difference between an expert and layman's knowledge on the subject. Overall, participants showed a weakening of the IOED with decreased ratings of understanding following their explanations; however, participants in the devices condition had higher ratings of understanding at Time 1 than those in the disorders condition and experienced a greater change in ratings. Additionally, participants in the devices condition had higher ratings for society's knowledge and a smaller gap between experts and laypeople than those in the disorders condition. The weakening of the IOED seen for mental disorders was only present when participants

were providing explanations; when participants were randomly assigned to either explain or list characteristics about a series of mental disorders, those in the description condition did not significantly differ in their Time 1 to Time 2 ratings of understanding. These findings suggest not only that description is not enough to weaken the IOED, but also that the strength of the illusion may be related to the extent to which individuals believe a concept is understood by experts, as disorders had a smaller change in ratings of understanding than devices.

Despite people's trust in others' knowledge being linked to the magnitude of the weakening of the IOED, experts are not immune to illusions about their own knowledge. Fisher and Keil (2015) asked participants to explain either three familiar topics or three unfamiliar topics in an IOED paradigm. In Experiment 1, the familiar topics were based on gender- and age-based topics that corresponded to the participant's reported gender and age; the unfamiliar topics were gender- and age-related topics from another demographic. Unsurprisingly, the weakening of the IOED was stronger for lower-education participants who were asked to explain more familiar topics. However, when college-educated participants were asked to explain either topics from within or outside their major, they unexpectedly experienced a greater breaking of the IOED for topics within their major. According to Fisher and Keil, "Formal expertise leads to illusions of understanding for previously mastered formal topics" (2015, p. 1261). These findings are similar to that of Lichtenstein and Fischhoff (1977), who also found that domain expertise is not necessarily related to improved calibration in that domain.

Although the domains in which the IOED are applicable are limited, weakening the IOED is not necessarily domain-dependent and can occur when asking participants to

explain a concept in a different domain. Roeder and Nelson (2015) found that asking participants to explain a topic in a domain unrelated to what was being rated for understanding (e.g., explaining "how an official is elected to the Nigerian House of Representatives" when rating understanding for how a helicopter flies) still resulted in a weakening of the IOED. This suggests that the act of explanations alone can cause participants to adopt a new criterion for what constitutes a high level of knowledge and adjust their assessments accordingly, resulting in potentially more calibrated responses.

Can the IOED Lead to Improved Performance?

As seen previously, simply thinking about a complex concept does not result in a reconsideration of one's knowledge on that topic (Zeveney & Marsh, 2016) – it is the act of explaining that forces the weakening of the IOED. How, then, could explanations help individuals to become better calibrated in their knowledge? It may be that the act of explaining material results in increased attention, and therefore better encoding, of the information at hand. Griffin, Wiley, and Thiede (2008) assigned introductory psychology students to read texts once, twice, or explain the text to themselves as they read. Participants' comprehension of the text and overall comprehension ability was then tested. Those who self-explained had the greatest calibration, and only lower comprehension ability participants showed improvements in monitoring accuracy when they read twice. There were no interactions between ability or working memory capacity with explanation, nor did the three groups differ in test performance. These findings suggest that the act of explanation may increase attention to cues when reading, leading to better metacognitive monitoring accuracy.

Even though explanations may improve calibration, there is no guarantee that this act should also increase performance. However, previous research suggests that the explaining may lead to better application of knowledge. Coleman, Brown, and Rivkin (1997) asked students to learn the information in a scientific text via explanation or summarization. Participants were then assigned to one of three tasks: studying for themselves, studying to teach others, or listening to someone else who either explained or summarized the task. Participants then completed a reading comprehension task, explained or summarized the topic of the text, and answered questions about the text that required near transfer (something related to an example in the text) and far transfer (required students to make inferences about the problem and apply in novel task). Despite having similar reading comprehension scores, students in the explanation group produced more ideas and showed better quality in their written responses and in their far transfer problem than those in the summarize condition. Additionally, those who were asked to study for themselves or study to teach others performed better than those who only heard the information. Those in the teaching condition performed better on the near and far transfer problems than the other two conditions. These results suggest that having to explain information to one's self or others may change how people encode or structure information when studying.

Although IOED procedures are thought to improve individual calibration, at the time of writing there has been little research done on whether or not weakening the IOED can lead to improved performance. However, the IOED has been shown to moderate people's beliefs and affect behavioral outcomes. Over three studies, Fernbach, Rogers, Fox, and Sloman (2013) asked participants to rate their understanding of certain political

positions. After doing so, participants provided either explanations or reasons for their ratings before re-rating their understanding. Participants who provided explanations experienced a significant drop in understanding at post-explanation ratings and became less extreme in their positions. Participants who provided reasons also experienced a slight decrease in understanding, but their position ratings did not change. The moderation in position rating for the explanation condition were reflective of more tangible changes as well. When participants were asked after their explanations or reasons if they wanted to donate to an advocacy group relevant to their interests, those in the explanation condition were less likely to donate, whereas those in the reasons condition were more likely to donate. Experiencing a weakening of the IOED not only causes individuals to reassess their level of perceived understanding, but also changes their behaviors with regard to that information. If people become less overconfident in their positions, they also are less likely to act upon those prior positions.

Some promising results suggest that IOED may improve learning. In their initial four studies, Rozenblit and Keil (2002) asked participants to read an expert's explanation of the items after re-rating their initial understanding. Participants then provided a new rating of their knowledge, keeping in mind what was considered an expert's level of understanding of the item. Notably, when given an expert's explanation of a concept, participants reported increases in their own knowledge about a topic. Separately, the act of retrieving the explanations alone may assist in increased performance. Previous research has found a robust effect of retrieval on improved long-term memory retention (Bjork, Dunlosky, & Kornell, 2013; Karpicke & Roediger, 2008; Roediger & Karpicke, 2006a; Roediger & Karpicke, 2006b; Roediger & Pyc, 2012). Although there are

important distinctions between generating explanations and other retrieval methods like testing, they have both been shown to similarly improve later memory for information (e.g., Larsen, Butler, & Roediger, 2013; Mulligan & Peterson, 2015). Importantly, McDaniel, Roediger, and McDermott (2007) found that completing a short-answer test, particularly when receiving immediate feedback, led to improved performance on a final test compared to reading or multiple-choice testing. Given the similarity between short-answer testing with immediate feedback and giving explanations followed by an expert's explanation, it is possible that participants who undergo the IOED paradigm will see comparable benefits in long-term retention.

By having people acknowledge their limitations on knowledge of a topic, as well as attempt to retrieve the information, before reading expert information, they may be more attentive to the new information. This is of particular interest for informing jurors about the fallibility of memory and creating jurors who are not only more skeptical of eyewitness information, but more sensitive to the factors that render eyewitness memory more or less credible. Given the relative ease with which modified versions of the IOED paradigm can be employed, this methodology may be a promising tool for combatting juror overconfidence about how memory works.

Cognitive Reflection May Moderate the Weakening of the IOED

Unquestionably, jurors are required to process and evaluate a large amount of material over the course of a trial. Although ideally jurors are relying on effortful processing, jurors (whose cognitive resources are depleted from trying to keep track of evidence, statements, testimony, and deciphering legal jargon) likely rely on heuristics and biases instead to make their decisions. Jurors who are overwhelmed by information

are more likely to rely on stereotypes – particularly those which are more congruent with criminality – when making judgments about guilt (e.g. Van Knippenberg, Dijksterhuis, & Vermeulen, 1999).

Previous literature suggests that a juror's preference for engaging in effortful thinking influences his or her decision making. Leippe, Eisenstadt, Rauch, and Seib (2004) found that jurors with moderately high preference for engaging in effortful thinking, as measured by the Need for Cognition scale (NFC; Cacioppo, Petty, & Kao, 1984), were more likely to convict when the case against the defendant was strong, but those with the lowest and highest NFC were less likely to convict. Separately, McAuliff and Kovera (2008) found that jurors higher in NFC were more likely to agree with the plaintiff in a civil case when the expert's testimony was internally valid, whereas low NFC jurors were not sensitive to the difference in validity. Thus, one would expect that those jurors who prefer to engage in effortful thinking will differentially respond to evidence presented at trial.

Previous research in the lab found NFC did not explain unique variance in juror decision making (Malavanti, 2014). Although NFC may indeed influence some juror decision making, it is not clear whether a preference for thinking alone will cause them to update and therefore re-calibrate their memory knowledge. However, the objective tendency to engage in reflection may influence whether a person is likely to experience the weakening of an IOED and therefore re-assess their knowledge. Fernbach, Sloman, St. Louis, and Shube (2013) found that those who are less likely to engage in cognitive reflection are likely to experience a greater change in ratings of understanding than those who engage in more cognitive reflection. In an IOED paradigm, participants were asked

to read four levels of explanations (no explanation, shallow explanation, intermediate explanation, and detailed explanation) for how each of four products worked. Next, participants rated the explanations on their level of understanding before completing the Cognitive Reflection Test (CRT; Frederick, 2005). The CRT assess an individual's preference for more deliberate, reflective reasoning. Participants who scored lower on the CRT felt they understood the shallow explanations more than the detailed and felt they received more understanding from the no detail explanation compared to high CRT participants (Fernbach, Sloman, et al., 2013). Additionally, lower CRT participants preferred products most when accompanied shallow explanations, whereas higher CRT participants' preferences increased with detail. Interestingly, the effects seen based on CRT scores only pertained to causal details; both high and low CRT groups showed an increase in understanding as non-causal detail increased. Together, these findings suggest that lower CRT individuals are more susceptible to illusions of understanding for causal details.

Unlike measures of cognitive preference like the NFC, the CRT is not based on self-report. This difference may explain why lower NFC participants showed no difference and higher NFC participants' level of understanding increased with both causal and non-causal detail (Fernbach, Sloman, et al., 2013). These findings suggest that NFC and similar measures may reflect a preference for detail in explanations, but level of cognitive reflection moderates the causal aspect associated with explanatory knowledge. Therefore, I expect that those individuals who are less likely to engage in reflective processing will experience a greater weakening of the IOED and show the greatest

improvement in their calibration on the memory task. However, those who have higher CRT scores will likely perform the best on any memory-related tasks.

Overview of Experiments

To my knowledge, the current research is the first to investigate memory knowledge in the context of the IOED paradigm, as well as one of the first to explore performance changes due to the weakening of the IOED. This investigation not only extends current experimental research exploring the relevant domains of the IOED, but also adds to the existing literature on calibration and learning. In Experiment 1, I examined whether perceived memory knowledge is susceptible to a weakening of the IOED. In Experiment 2, I extended these findings to explore if the IOED paradigm can assist participants in correctly assessing their own knowledge and if undergoing a weakening of the IOED potentially improves later performance. Finally, I extended these findings to the courtroom using an online juror decision making paradigm. In Experiment 3, I investigated the utility of the IOED paradigm as a tool for reducing juror over-reliance on memory by examining its effectiveness in an applied setting. Because the general public remains misinformed or under-informed about how memory works, these studies provide an initial look into a promising means for educating prospective jurors about how memory works, and the results from this project can be applied to other learning contexts in addition to juror decision making.

CHAPTER TWO

Experiment One

Overview

In Experiment 1, I investigated whether explaining weakens the illusion of explanatory depth (IOED) for memory in a similar way as it does mechanical devices, mental disorders, and natural phenomenon. Participants underwent an IOED paradigm for either memory phenomena (e.g., "How are memories stored") or devices (e.g., "How a greenhouse maintains temperature") and rated their level of understanding for these concepts a total of four times.

Hypotheses

I had four central hypotheses for this experiment:

1. Overall, participants would show decreases in their ratings of perceived understanding following their explanations (R2) and following the reassessment of their understanding after the expert's explanations (R3).
2. Overall, participants would show an increase in their rated understanding following the expert's explanations (R4) compared to their initial assessment of their knowledge (R1).
3. Participants in the memory condition would show a pattern of responses similar to that of the devices condition; however, the weakening of the IOED would be greater for devices.

4. Participants in the memory condition would believe that the gap between an expert's and layperson's knowledge is greater than those in the devices condition.

Method

Participants

Study participants ($N = 100$) were recruited from Amazon's Mechanical Turk (MTurk). Only United States-based MTurk workers with at least 500 approved Human Intelligence Tasks (HIT) and a HIT approval rate of at least 90% were permitted to take part in this study.[1] Participants received $1.00 for their participation. Participants were predominantly White and nearly equally male and female (76.0% White; 51.0% Female). Additionally, participants were nearly equally split among conservative-leaning and liberal-leaning ends of the political spectrum (39.0% leaning to very conservative; 43.0% leaning to very liberal). A plurality had achieved a four-year college degree (34.0%), with the next largest group having achieved at least some college (29.0%). Most participants estimated their household income between $50,000-$59,999 (21.0%), with the next-largest group estimating their income at $20,000-$29,999 (20.0%). A majority of participants completed the experiment during the afternoon (12:00PM-5:59PM = 41.0%). Many participants claimed they had slept for 8 hours the night before the study (29.0%),

[1] According to Amazon Mechanical Turk, a Human Intelligence Task (HIT) is "a single, self-contained task that a Worker can work on, submit an answer, and collect a reward for completing" (Amazon Mechanical Turk, 2018, n.p.). The individual who created the HIT (known as a Requester) determines whether to approve or reject a Worker's work. The criteria for an approval vary by Requester; in these experiments, workers were approved provided they were at least 18 years old, they consented to participate, and they provided the correct code (received at the end of the study). Although there are various recommendations for setting HIT approval criteria, Workers with higher approval ratings often provide higher quality data (see Peer, Vosgerau, & Acquisti, 2014). I report our specific criteria here in line with MTurk best practices (see Burhmester, Talaifar, & Gosling, 2018).

with most claiming they needed either 8 (34.0%) or 7 hours (26.0%) of sleep on average. All participants indicated they were at least 18 years of age (M_{Age} = 35.88; *SD* = 10.78; Range = 19-63), and nearly all indicated they were eligible to vote in the United States (98.0% Eligible to vote); these final two questions served as proxies for participants' eligibility for juror selection for courts in the United States (United States Courts, 2017).

Materials

IOED questionnaire. Following the methodology from Rozenblit and Keil (2002), I created a questionnaire designed to elicit an illusion of explanatory depth (IOED) for the domain of memory (see Appendix B). I used items from a previous IOED questionnaire for devices (Rozenblit & Keil, 2002) to serve as the comparison condition. All participants read the instructions on how to rate their level of understanding taken also from Rozenblit and Keil (2002). These instructions included examples of an explanation at a low, medium, and high level of knowledge taken from Fernbach, Rogers, and colleagues (2013). Participants were instructed to use these examples to help them assess their level of knowledge on the subject. The instruction text was at a 10th grade level according to the Flesch-Kincaid Grade Level test; the texts for memory and devices were around a 9th grade level (M_{Memory} = 9.48; $M_{Devices}$ = 9.08). These questionnaires were piloted (*N* = 24) to ensure that the materials were clear and that participants were responding to the items as intended.

Dependent Measures

Level of understanding. Participants rated their level of understanding on a scale from 0-100% (0 = No understanding, 100 = Complete understanding).

Society rating. Participants rated how much they believed society understands each of the phenomena on a 0-100% scale (0 = Society understands 0% of this concept, 50 = Society understands 50% of this concept, 100 = Society understands 100% of this concept). This item was adapted from Zeveney and Marsh (2016).

Gap rating. Using a scale from 0-100% (0 = 0% difference, 50 = 50% difference 100 = 100% difference), participants rated the size of the difference between an expert and a layperson's knowledge on each of the phenomena. This item was also adapted from Zeveney and Marsh (2016).

Procedure

Participants were told that they would be completing an online questionnaire about knowledge. All participants were required to indicate their informed consent before being permitted to participate in the study. Participants who consented to the research then completed basic demographic information, including their age, gender, race, political affiliation, highest level of education, socioeconomic status, as well as items asking about their sleep habits.

At the start of the study, participants were assigned a subject number to safeguard their identity and were referred to by that subject number throughout the study and analysis to ensure their privacy and confidentiality. Participants were randomly assigned to one of two treatment levels before the study session: Devices (n = 50) or Memory (n =

50). After reading the instructions, all participants read the five concepts from their assigned domain and rated their level of perceived understanding on those topics (R1; see Figure 2.1). Following this, participants were asked to explain the concepts and re-rate their understanding (R2) before being shown explanations from experts on their assigned concepts. Finally, participants were asked to re-assess their initial understanding of the concepts (R3) and rate how much they learned from the expert explanations (R4). At the end of the experimental session, participants completed the society and gap ratings for the five concepts before being debriefed on the purpose of the study and given the contact information of the IRB and experimenters.

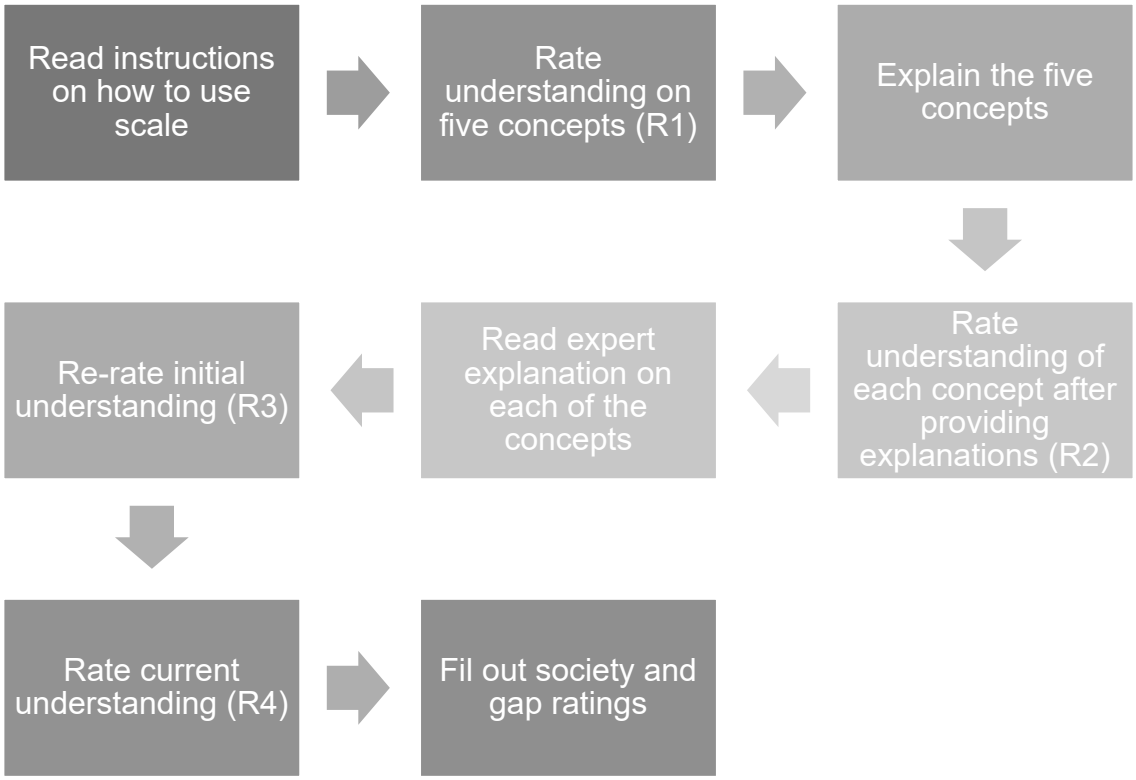

Figure 2.1. Methodology for Experiment 1.

Statistical Analyses

Data Analysis

To assess participants' changes in levels of understanding, I analyzed the data with a linear mixed-effects model with condition (Devices or Memory), rating (1, 2, 3, or 4), and their interaction as fixed effects and participants and concepts as random intercepts. Devices and R1 served as baseline conditions. Participants' reported level of understanding were mean-centered by participant. To assess participants' society and gap ratings, I analyzed the data with a linear model with condition (Devices or Memory) as a fixed effect and participants and concepts as random intercepts.

Results

Level of Understanding

Participants differed in baseline ratings of claimed understanding between conditions, B = -6.26, SE = 1.79, β = -0.14, $p < .001$, 95% CI [-9.77, -2.76] Participants were less likely to indicate understanding of memory concepts compared to devices at their initial rating. Participants' ratings did change from baseline as they progressed through the task. Ratings of understanding for those in the devices condition decreased significantly from the first rating to the second (B = -8.35, SE = 1.64, β = -0.41, $p < .001$, 95% CI [-11.57, -5.13]), decreased significantly from the first to third (B = -13.54, SE = 1.64, β = -0.30, $p < .001$, 95% CI [-16.76, -10.32]), and increased significantly from the first to final rating (B = 12.98, SE = 1.64, β = 0.64, $p < .001$, 95% CI [9.76, 16.20]) However, these main effects were qualified by significant condition by rating interactions (see Figure 2.2). Participants in the Memory condition reported higher ratings of

understanding than those in the Devices condition at the third (B = 10.16, SE = 2.32, β = 0.50, p < .001, 95% CI [5.61, 14.72]) and fourth ratings (B = 11.92, SE = 2.32, β = 0.26, p < .001, 95% CI [7.36, 16.48]). There was no significant effect of memory condition compared to devices at the second rating (B = 2.97, SE = 2.32, β = 0.06, p = .202, 95% CI [-1.59, 7.52]). Planned comparisons with Bonferroni correction revealed although those in the Devices condition showed significant decreases in understanding from the first and second rating and from the second to third rating (ps < .01), those in the Memory condition only showed a significant decrease between the first and second rating (p = .007). Both groups showed a significant increase in perceived understanding from the third to fourth rating (ps < .001).

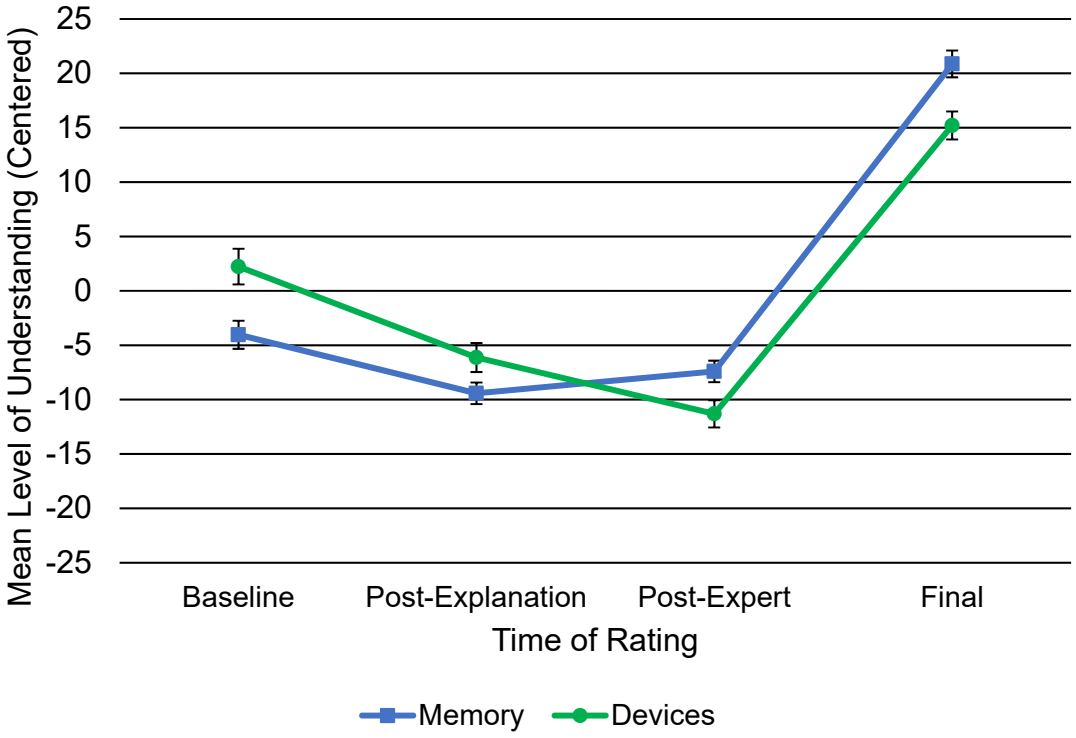

Figure 2.2. Mean level of perceived understanding changed over time by condition. Note: Baseline = R1; Post-Explanation = R2; Post-Expert = R3; Final = R4.

Society and Gap Ratings

Participants did not differ in their assessments of how much society understands the device or memory concepts, B = -1.90, SE = 4.42, β = -0.04, p = .669, 95% CI [-10.56, 6.77]. Mean ratings for devices (M = 38.07, SEM = 1.58) and memory concepts (M = 36.17, SEM = 1.59) overall were fairly similar. The groups were also comparable in their assessments of the gap between expert and layperson knowledge, B = 0.72, SE = 3.68, β = 0.02, p = .846, 95% CI [-6.50, 7.94]. Again, mean ratings for devices (M = 68.98, SEM = 1.42) and memory concepts (M = 69.69, SEM = 1.41) were similar. As seen in Table 2.1, society ratings were positively related to all four uncentered ratings of understanding, whereas gap ratings were only significantly correlated with R4. Overall society and gap ratings were not significantly related. Dividing participants by condition did not alter my reported pattern of findings for society ratings; however, different patterns emerged for the gap ratings. For those in the memory condition, gap ratings were significantly negatively related to all ratings of understanding ($r_{\text{Memory R1}}$ = -.16, $r_{\text{Memory R2}}$ = -.16, $r_{\text{Memory R3}}$ = -.13, $r_{\text{Memory R4}}$ = -.21; ps < .05). For participants in the devices condition, gap ratings were only significantly negatively correlated with R1 ($r_{\text{Devices R1}}$ = -.14, p = .019).

Table 2.1

Correlations Among Ratings of Understanding and Society and Gap Ratings

Measures	1	2	3	4	5
1. R1	1				
2. R2	.71**	1			
3. R3	.59**	.77**	1		
4. R4	.49**	.54**	.62**	1	
5. SR	.51**	.64**	.76**	.60**	1
6. GR	-.003	-.003	.002	.14*	-.05

Note. SR = Society rating, GR = Gap rating, * $p < .01$, ** $p < .001$.

Discussion

Experiment 1 provided support for memory as a domain susceptible to illusions of explanatory depth. As expected, participants in both the Devices and Memory conditions showed a decrease in ratings of perceived understanding from baseline after they were challenged to provide explanations of the target concepts and a reported increase in understanding after reading the expert explanations, replicating previous work by Rozenblit and Keil (2002). Together, these findings bolster the first and second hypotheses regarding the overall pattern of results and provide initial evidence of the utility of the IOED task for challenging people's illusions of knowledge about memory.

However, the two conditions did differ throughout the task, as reflected in the condition by rating interactions observed in Experiment 1. Although the initial ratings of claimed understanding for devices was higher than that for memory, participants in the Memory condition reported higher ratings of understanding than those in the Devices condition at the third and fourth ratings. The differences at these ratings may be explained by the fact that those in the memory condition did not significantly change their ratings from R2 to R3. Whereas those who completed the task for devices

demonstrated the expected decreases from R1 to R2 to R3, Memory participants reported a decrease in ratings only after providing explanations and did not reassess their understanding as lower at R3. It was from this higher reported understanding that they assessed their new understanding at R4.

Although the overall results support my third hypothesis that the weakening of the IOED would be greater for those in the Devices condition, those in the Memory condition estimated their knowledge higher at R3. However, this pattern is consistent with previous work. Rozenblit and Keil (2002) observed that people showed a steeper decline for their level of understanding devices than for natural phenomena. Therefore, it is probable that much like other natural phenomena, illusions of understanding for memory are easier to expose and weaken compared to those for devices. As the number of visible compared to hidden parts strongly contributes to people's illusions of explanatory depth (Rozenblit & Keil, 2002), devices likely encourage more overconfidence than memory – a mostly hidden phenomenon – does. Therefore, when challenged, people are more likely to report a lack of understanding for natural phenomena, including memory, than for devices. Supporting this idea is the fact that those in the Memory condition rated their level of understanding lower than those in the Devices condition at both R1 and R2. Participants rating memory concepts were quicker to estimate their knowledge as low than those rating devices, and thus were not as compelled to re-evaluate their level of understanding at R3. Further research is needed to completely tease out the mechanisms behind these differences.

In contrast to the expectations for the ratings of perceived understanding, I did not find any support for the fourth hypothesis regarding participants' society and gap ratings.

Participants believed that society's understanding for the devices and memory concepts were similarly low. Additionally, participants believed that the gap between layperson and expert knowledge was similarly large for devices or memory concepts. Exploratory analyses revealed that overall society ratings were positively related to reported level of understanding regardless of rating instance, and dividing participants by condition did not change this pattern. Thus, as participants believed they understood more about the topic, their ratings for what society understood also increased. This relationship is consistent with previous work on the "curse of knowledge," or the idea that people often assume that others know what they do (Camerer, Loewenstein, & Weber, 1989). Because this relationship persisted across ratings, these society ratings likely reflect participants' overall perceptions of what they believe society knows, rather than reflecting their individual changes in level of understanding due to some aspect of the rating task, i.e., feelings of ignorance after explaining. These findings replicate those of Zeveney and Marsh (2016), who saw a similar relationship between society ratings and ratings of understanding for mental disorders and devices. Further, the observed relationship between participant ratings and society ratings for memory knowledge echoes that found by Malavanti, Terrell, Dasse, and Weaver (2014), who observed that attorneys – who often expect laypersons to understand how memory works – in fact have a better understanding of memory than the general public.

Overall gap ratings, on the other hand, were associated only with participants' reported understanding at the end of the task but dividing participants by conditions altered the direction of these effects. For those in the memory condition, gap ratings were negatively associated with all rating instances. Going through the task did not appear to

have altered their belief that as their own personal understanding decreased, the knowledge gap between laypersons and experts increased. These findings echo those observed for the mental disorders (Zeveney & Marsh, 2016). However, unlike those in the Devices condition in Zeveney and Marsh's (2016) study, the initial ratings of understanding for those in my Devices condition were negatively associated rather than unassociated with gap ratings. Because only their initial ratings were related to their perception of the size of the gap, it may be that these participants were affected by their experience during the task. As they experienced a continued decrease in reported understanding, those in the devices condition may have stopped relying on their personal understanding as a cue for the size of the knowledge gap. It is important to note that participants made their society and gap ratings at end of the task, and therefore their perceptions of what society or experts know about how memory or devices work may reflect information they learned during the task.

Generally, the results of Experiment 1 indicate that memory is a viable domain for the IOED task. Memory knowledge does appear susceptible to an illusion of understanding, adding to a growing list of domains where people are prone to conflate shallow knowledge with a deeper one. Encouragingly, people do appear to feel their knowledge has increased after receiving expert information, suggesting that weakening the illusion may improve not only people's assessment of their understanding, but also later learning. However, whether these changes in reported understanding reflect actual improvements in metacognition or learning remains to be investigated.

CHAPTER THREE

Experiment Two

Overview

In Experiment 2, I explored the utility of the IOED for learning. I investigated whether 1) the IOED paradigm improves performance or calibration of knowledge on a later assessment compared to a less intensive task, i.e., listing what they know about the topic, and 2) whether any improvement in performance or calibration is domain- or item-specific. Participants rated their understanding of how memory works before either explaining or listing information about their assigned sub-concepts. After re-rating their understanding, participants completed a memory questionnaire and estimated the probability that they answered the questions correctly.

Hypotheses

I had three main hypotheses for this experiment:

1. Participants who provided explanations would experience a greater change in ratings of claimed understanding than those who listed.
2. Those who underwent the IOED task would be more calibrated in their assessments of their performance compared to those who listed.
3. The weakening of the IOED would be domain agnostic; that is, those who completed the IOED for devices would show a similar change in ratings of understanding and metamemory accuracy as those who completed the IOED for memory.

Method

Participants

I recruited 221 participants from MTurk for this study. Participants recruited directly via MTurk were required to meet the same HIT requirements as those in Experiment 1; however, a portion of the MTurk sample was recruited using the Turk Prime toolkit (Litman, Robinson, & Abberbock, 2016). These workers were required to have at least 1000 approved HITs and a HIT approval rate of at least 98%. All participants were paid $1.50 for their work. Forty-one participants were removed from the study due to suspicious, plagiarized, or nonsensical responses. This left a final sample of 180 participants (77.2% White; 55.6% Male). Only one participant indicated they had not received at least a high school diploma or equivalent; 90.0% of participants noted they had completed at least some college, with nearly two-thirds indicating they had obtained at least a four-year degree (62.2%). Participants were 33.9% Conservative-leaning, 21.1% Moderate, and 41.1% Liberal-leaning. A plurality of participants (13.9%) estimated their household income at between $40,000-$49,999; the next-largest group estimated their household income at $70,000-$79,999 (12.8%). Additionally, most participants (43.3%) indicated they completed the survey during the afternoon (12:00PM – 5:59PM). The largest group of participants (22.8%) estimated they had 8 hours of sleep the night before, and 31.7% believed they needed 8 hours of sleep per night. All participants were at least 18 years old (M_{Age} = 36.16; *SD* = 11.73; Range = 18-72) and nearly all indicated they were eligible to vote in the United States (98.9% Eligible to vote).

Materials

IOED questionnaire. I used the same IOED questionnaire for devices that was employed in Experiment 1. The IOED questionnaire for memory was modified such that participants rated their understanding on concepts covered in the expert information and memory questionnaire (see Appendix C).

Jury instructions. These instructions were a modified version of the *Henderson* instructions, established in 2012 by the New Jersey Supreme Court (*New Jersey v. Larry R. Henderson*, 2011). Because there was no case accompanying this experiment, I omitted any language referring to presented evidence or testimony. The instructions scored a 14.5 on the Flesch-Kincaid Grade Level test.

Memory questionnaire. I created a 16-item memory questionnaire by pulling items from the 30-item memory questionnaire from Kassin and colleagues (2001) and the six items from Simons and Chabris (2011, 2012). Both sets of items covered several topics related to eyewitness memory. For my purposes, I included 8 items covered in the jury instructions and 8 items not covered in the experiment; the memory questionnaire is available upon request. Participants read a statement, such as *Very high levels of stress impair the accuracy of eyewitness testimony*, and indicated whether or not they believed the statement was supported by the current scientific literature (0 = No, 1 = Yes). Participants also noted how confident they were that they answered the item correctly. The memory questionnaire was piloted before being employed in this experiment. Participants in the pilot ($N = 30$) indicated an average of 5.57 of the 8 critical memory concepts were present in the modified jury instructions ($SEM = 0.33$), and all critical

memory concepts were noted as present in the jury instructions, with percentages ranging from 93.3% ("Stress and memory for an event") to 46.7% ("Storing of memories" and "Level of confidence and the accuracy of a memory").

Cognitive Reflection Test. The original Cognitive Reflection Test (CRT; Frederick, 2005) is a three-item measure intended to assess an individual's preference for reflective reasoning. The items are short word problems that offer a seemingly simple incorrect response (e.g. *A bat and a ball cost $1.10 in total. The bat costs $1.00 more than the ball. How much does the ball cost?*) but can be easily answered after reflection. The CRT has shown good predictive validity with measures of delayed gratification and gambling. Further, the CRT has moderate positive relationships with measures of intelligence, such as the Need for Cognition scale (NFC; Cacioppo, Petty, & Kao, 1984), but accounts for distinct variance in decision-making tasks (Frederick, 2005). Due to the popularity of this measure, I used an extension of the CRT developed by Toplak, West, and Stanovich (2013). This seven-item CRT measure is moderately positively related to the original and has shown good reliability ($\alpha = .72$) and similar predictive ability on rational thinking tasks (Toplak, West, & Stanovich, 2013). The seven-item CRT showed adequate reliability in my sample ($\alpha = .66$).

Dependent Measures

Level of understanding. Participants rated their level of understanding on the same 0-100% scale as in Experiment 1.

Confidence score. After responding to each of the 16 items on the memory questionnaire, participants indicated the probability that they answered the question correctly. As the probability of answering a question correctly by guessing was 50%, participants rated their confidence on a 50-100% scale (50 = 50% confidence, 100 = 100% confidence).

Accuracy. Participants' actual performance on the memory questionnaire was scored dichotomously (0 = Incorrect, 1 = Correct).

Procedure

Before the study, participants were randomly assigned to one of three treatment levels (Condition: Explain-Memory or Explain-Devices or List-Memory). Participants filled out the same demographic information as in Experiment 1 at the start of the study. After filling out demographic data, all participants were instructed on how to fill out the levels of understanding ratings before rating their knowledge on eight memory items drawn from the jury instructions (R1; see Figure 3.1). Afterward, participants in the IOED conditions were asked to explain four of their assigned concepts (Memory or Devices), and participants in the listing condition were asked to list, using bullet points, what they knew about four of the memory concepts in lieu of the explanation task. After providing their explanations, participants were asked to re-rate their levels of understanding of the eight memory concepts (R2).

Next, all participants were asked to read the modified jury instructions. After reading the instructions, participants rated their level of understanding a final time before completing the memory questionnaire. Participants were not told about the questionnaire

before receiving it. After finishing the questionnaire, participants completed the multiple-choice version of the seven-item Cognitive Reflection Test (CRT; Sirota & Juanchich, 2018; Toplak, West, & Stanovich, 2013) before being debriefed on the nature of the study.

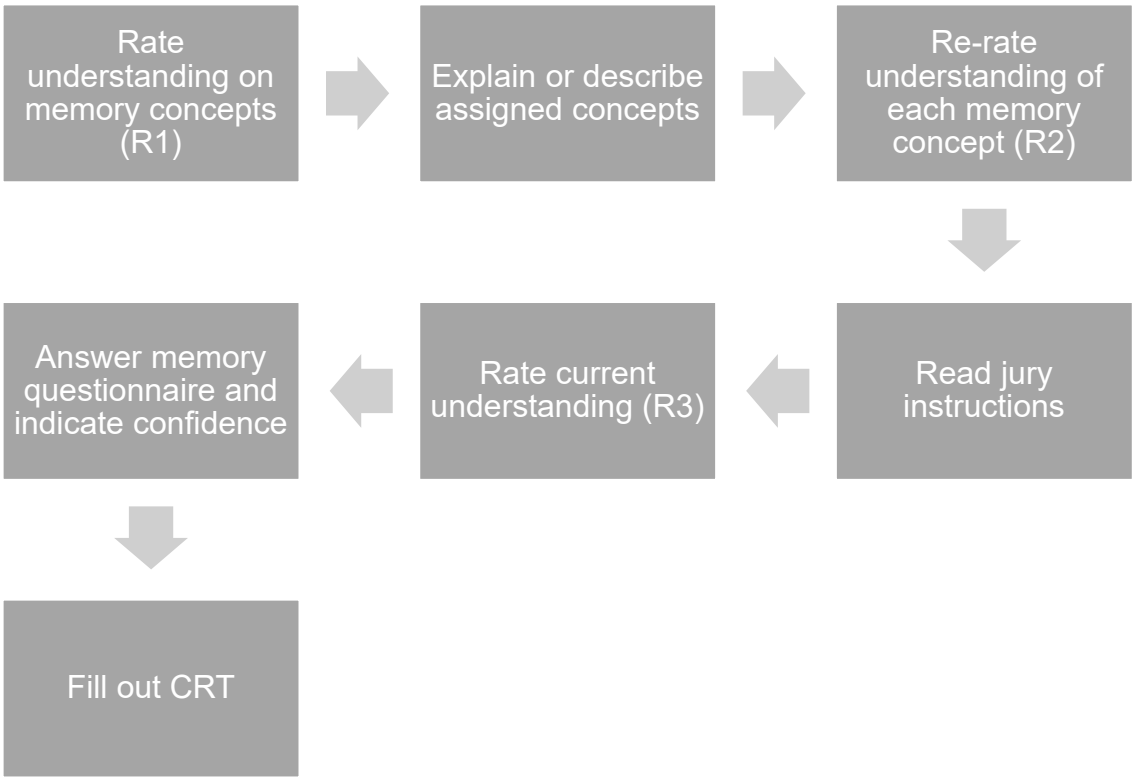

Figure 3.1. Methodology for Experiment 2.

Results

Level of Understanding

To assess participants' changes in levels of understanding, I analyzed the data with a linear mixed-effects model with condition (Condition: Explain-Memory, Explain-Devices, or List-Memory), Rating (1, 2, or 3), and their interaction as fixed effects and participants and concepts as random intercepts. Explain-Memory and R1 served as reference groups. Participants' reported level of understanding were mean-centered by participant.

Participants did not differ by condition in their initial ratings of understanding for the memory items. Neither those in the Explain-Devices nor those in the List-Memory condition significantly differed from those in the Explain-Memory condition, $B = -0.36$, $SE = 0.64$, $\beta = -0.02$, $p = .720$, 95% CI [-2.32, 1.60]; and $B = 0.92$, $SE = 0.66$, $\beta = 0.05$, $p = .369$, 95% CI [-1.08, 2.92], respectively. Additionally, participants did differ in ratings over time. Ratings of understanding for those in the Explain-Memory condition were significantly lower than baseline at the second rating, $B = -5.24$, $SE = 0.97$, $\beta = -0.26$, $p < .001$, 95% CI [-7.15, -3.33], and were significantly increased from baseline at the third rating following the expert information, $B = 6.40$, $SE = 0.97$, $\beta = 0.32$, $p < .001$, 95% CI [4.49, 8.31]. These main effects were qualified by a significant condition by time interaction, $B = -2.72$, $SE = 1.38$, $\beta = -0.13$, $p = .050$, 95% CI [-5.42, -0.01], such that participants in the List-Memory condition had decreased ratings of understanding at the second rating compared to those in the Explain-Memory condition (Figure 3.2). There were no other significant interactions. Adding CRT scores to the model did not alter the reported pattern of results, and propensity for cognitive reflection did not significantly

account for any variance in ratings of understanding, $B < .001$, $SE = 0.12$, $\beta < .001$, $p > .999$, 95% CI [-0.23, 0.23].

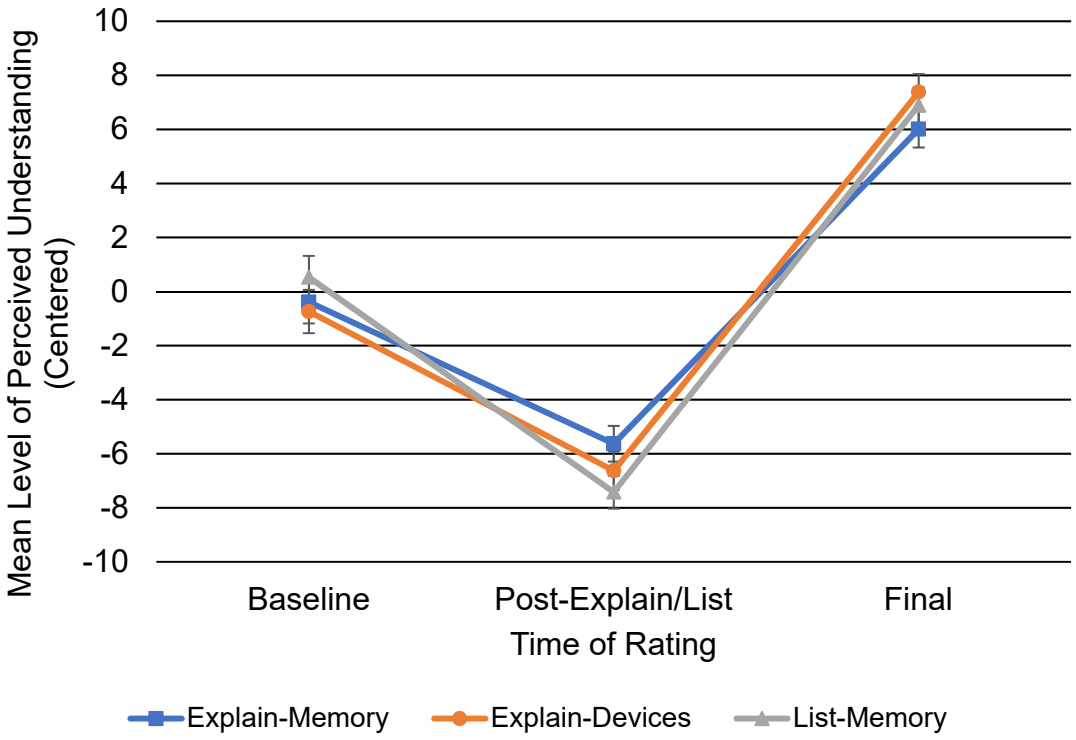

Figure 3.2. Mean perceived understanding changed similarly over time for each group. Note: Baseline = R1; Post-Explain/List = R2; Final = R3.

Performance

On average, participants answered 11.27 of the 16 questions correctly (*SEM* = 0.17; Range = 2-16). To explore any possible condition effects on performance, I conducted a linear regression with condition (Condition: Explain-Memory, Explain-Devices, or List-Memory) as a fixed effect and participants as random intercepts. Again, Explain-Memory served as the reference group. Participants' overall score on the memory questionnaire was similar across conditions ($M_{\text{IOED-M}} = 11.27$, $SEM = 0.30$; $M_{\text{IOED-D}} = 11.18$, $SEM = 0.29$; $M_{\text{LIST}} = 11.38$, $SEM = 0.31$). Participants performed

similarly between the Explain-Memory group and the Explain-Devices group, B = -0.10, SE = 0.42, β = -0.02, p = .820, 95% CI [-0.93, 0.73], and between the Explain-Memory group and the List-Memory group, B = 0.11, SE = 0.43, β = 0.02, p = .803, 95% CI [-0.74, 0.95]. Adding CRT scores to the model, however, did account for significant variance in overall performance, B = 0.45, SE = 0.07, β = 0.44, p < .001, 95% CI [0.31, 0.60]. Participants higher in cognitive reflection scored higher on the memory questionnaire than those lower in cognitive reflection.

Metacognition

Participants were fairly confident in their answers to the questionnaire (M = 83.17, SEM = 0.63).[1] Overall confidence ratings on the rated memory concepts were only related to participants' final ratings of understanding, r_{R3} = .15, p = .041; confidence and initial ratings, r_{R1} = -.05, p = .505, and second ratings, r_{R2} = -.052, p = .515, were not significantly correlated. To assess the utility of the instructions and the memory concepts they introduced on participants' performance, I ran a series of mixed-effects logistic regression models that included accuracy as the dependent variable, confidence as a fixed effect, participants and items as random intercepts, and confidence judgments by participant as a random slope (see Table 3.1). Participants' confidence in their individual judgments were mean-centered by participant, and the models predicted correct responses.

Confidence positively predicted memory performance, B = 0.05, SE = 0.01, p < .001, 95% CI [0.03, 0.07], OR = 1.05. Participants were slightly more likely to get

[1] Reported mean is uncentered for ease of interpretation.

questions correct as their confidence increased. To assess the effect of cognitive reflection on participants' metacognitive accuracy, I added CRT scores and a CRT by confidence interaction into the regression model. Adding CRT scores to the model accounted for significant additional variance above that of confidence ratings, $B = 0.31$, $SE = 0.05$, $p < .001$, 95% CI [0.21, 0.42], $OR = 1.37$. Participants higher in CRT were more likely to answer correctly on the memory questionnaire. Adding the covariate removed the previously reported finding for confidence on accuracy; with CRT scores in the model, confidence no longer significantly predicted performance, $B = 0.02$, $SE = 0.02$, $p = .294$, 95% CI [-0.02, 0.06], $OR = 1.02$. The CRT by confidence interaction approached significance, $B = 0.01$, $SE = 0.005$, $p = .072$, 95% CI [-0.001, 0.02], $OR = 1.01$.

I next ran a separate model with item type (Old or New) and an item type by confidence interaction added as fixed effects to assess how the memory concepts previously rated during the task influenced memory performance. New items served as the reference group. Participants were more likely to get questions correct when the items covered memory concepts that had been previously seen, $B = 2.49$, $SE = 0.63$, $p < .001$, 95% CI [1.26, 3.73]. Participants were around twelve times more likely to get old items correct than new items, $OR = 12.11$. Confidence remained a significant positive predictor of accuracy, $B = 0.03$, $SE = 0.01$, $p = .003$, 95% CI [0.01, 0.05], $OR = 1.03$, and the item by confidence interaction was significant, $B = 0.06$, $SE = 0.005$, $p < .001$, 95% CI [0.05, 0.07], $OR = 1.06$. Participants' confidence was a stronger predictor of accuracy for old compared to new items.

Table 3.1

Regression Coefficients for Models Predicting Accuracy and Metacognition

Variables	Model 1		Model 2		Model 3		Model 4	
	B	*SE*	*B*	*SE*	*B*	*SE*	*B*	*SE*
Intercept	1.83	0.46***	0.67	0.49	0.55	0.46	1.93	0.49***
Confidence	0.05	0.01***	0.02	0.02	0.03	0.01**	0.09	0.02***
CRT			0.31	0.05***				
CRT*Confidence			0.01	0.01†				
Old Item					2.49	0.63***		
Old Item*Confidence					0.06	0.01***		
Explain-Devices							-0.20	0.31
List-Memory							-0.07	0.32
Explain-Devices*Confidence							-0.06	0.03*
List-Memory*Confidence							-0.05	0.03*

Note. Models predict participants' likelihood of getting a correct response on the questionnaire. For Model 3, New Item is the reference group; for Model 4, Explain-Memory is the reference group. Confidence is mean-centered by participant. † $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$.

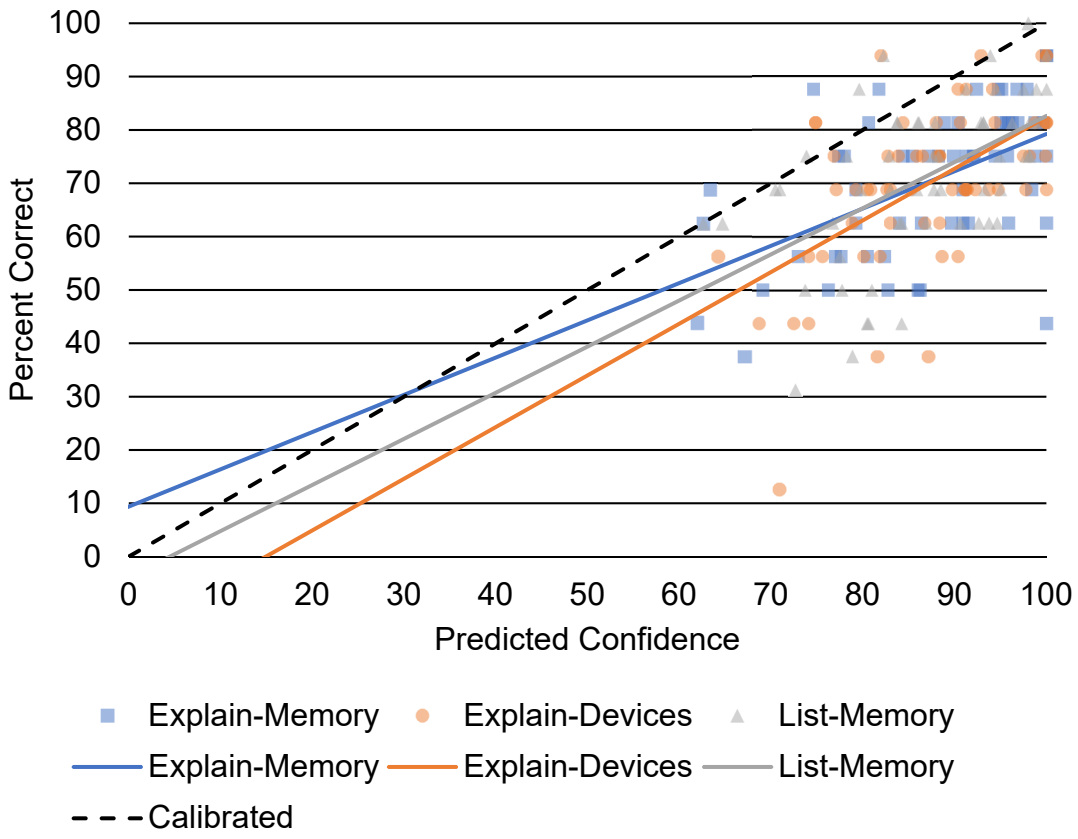

Figure 3.3. Simple calibration curves by condition for Experiment 2. The black dashed line represents perfect calibration, such that overconfidence is represented by the area below the line, and underconfidence by the area above the line. Although participants' predicted confidence was rated between 50-100%, the regression lines have been extended here for ease of comparison.

To investigate any effect of group on metacognition, I ran a model with condition (Explain-Memory, Explain-Devices, or List-Memory) and the condition-confidence interaction added to the base model as fixed effects. The Explain-Memory condition once again served as the reference group. Confidence ratings alone continued to positively predict accuracy even with these effects added to the model, $B = 0.09$, $SE = 0.02$, $p < .001$, 95% CI [0.06, 0.13], $OR = 1.10$. Completing the IOED task for devices or listing information about the memory concepts did not have a significant effect on participants' performance, $B = -0.20$, $SE = 0.31$, $p = .512$, 95% CI [-0.81, 0.13], $OR =$

0.82, and B = -0.07, SE = 0.32, p = .820, 95% CI [-0.69, 0.55], OR = 0.93, respectively. However, these findings were qualified by two significant condition by confidence interactions. Confidence was a weaker predictor of accuracy for both those who completed the Explain-Devices and List-Memory tasks compared to those in the Explain-Memory condition, B = -0.06, SE = 0.03, p = .015, 95% CI [-0.11, -0.01], OR = 0.94, and B = -0.05, SE = 0.03, p = .033, 95% CI [-0.10, -0.004], OR = 0.95. See Figure 3.3 for an aggregate representation of confidence on performance by group.

Discussion

As anticipated, participants in the two IOED conditions demonstrated the expected changes in ratings of understanding for the memory concepts. Regardless of the whether they explained memory concepts or devices, participants in both conditions showed decreases in their perceived understanding of memory after providing explanations and increases after reading expert information. These findings replicate my results from Experiment 1 that memory knowledge is susceptible to illusions of understanding; more importantly, these results provide further support that weakening of the IOED is not a domain-specific process (Roeder & Nelson, 2015). Engaging in the retrieval processes necessary for explanation, irrespective of what one is going to explain, appears to inform self-reported levels of understanding.

In contrast to the expectations stated in my first hypothesis, I also observed similar changes in ratings of understanding for those in the List-Memory and the Explain-Memory conditions. Despite starting at a similar baseline level of perceived understanding, those who listed information about the memory concepts showed a steeper decrease than those who explained at R2. These findings were surprising given that

previous literature on the illusion of explanatory depth has employed comparable conditions as a type of control where participants' level of claimed understanding remains stable across ratings (see Fernbach, Rogers, et al., 2013; Zeveney & Marsh, 2016). It is important to note that, in contrast to the aggregate analyses usually used in the IOED literature, e.g., repeated-measures ANOVA, the data in Experiment 2 were analyzed at the individual level. However, it is unclear why listing information about the memory concepts caused a greater decrease in ratings of perceived understanding than explaining. Even if participants in the listing condition were engaging in more retrieval or explanatory processes, their changes in ratings should have been akin to those who provided explanations. It is possible that the interaction observed at R2 was spurious or reflected the idiosyncrasies of those in the List-Memory condition, particularly in light of the fact that participants did not differ in their ratings of understanding after reading the expert information.

As all participants showed similar reported levels of understanding at the end of the task, it is not surprising that participants did not differ on the questionnaire regardless of condition, especially given that differences due to study condition are often only observable after a delay (see Roediger & Karpicke, 2006b; Van Den Broek, Segers, Takashima, & Verhoeven, 2014). However, investigating participants' metacognitive accuracy did reveal differences between the Explain-Memory and other two groups. As expected, participants' level of confidence in their responses was predictive of their accuracy; as participants were more confident in their retrieved answer, they were more likely to be correct. This is in line with previous metacognition research that generally finds higher confidence in retrieved answers predict accuracy on general knowledge tests

(e.g., Butterfield & Metcalfe, 2006; Perfect, Watson, & Wagstaff, 1993). As participants were actively trying to retrieve the correct response during the questionnaire, their feelings of confidence likely reflected their use of cues such as how easily they retrieved their answers and the amount of evidence for or against their chosen response (Costermans, Lories, & Ansay, 1992; Van Zandt, 2000). Thus, how more prospective measures of metacognition, or those taken before test, would compare to this more retrospective measure remains to be investigated. Although condition alone did not have any relationship with participants' likelihood to correctly answer the items on the questionnaire, confidence did interact with condition. For those in the Explain-Devices and List-Memory conditions, confidence was a weaker predictor of confidence than those in the Explain-Memory condition. These findings only offer partial support to my second hypothesis that those who explained would show better metacognitive accuracy than those who did not, as the results do not support a domain agnostic benefit to metacognition.

There is no clear reason why this should be the case, given participants' similar patterns of ratings of perceived understanding and overall scores on the questionnaire. It is possible that engaging in the explanation task for memory rather than devices increased participants' attention to the memory concepts and explanations provided in the jury instructions. Previous work has shown that participants' do attend more closely to feedback after they make high confidence errors or low confidence correct responses (Butterfield & Metcalfe, 2006). Thus, receiving feedback about the memory concepts after having overestimated their level of memory knowledge may have captured more of participants' attention in the Explain-Memory condition than the Explain-Devices

condition. Participants who suddenly explained devices, on the other hand, may have been thrown off by the change in topic and thus less aware of the importance of the memory concepts during the task.

However, this does not provide a completely adequate explanation for our observed results. All participants showed an increased likelihood to correctly answer previously-rated items, and participants' confidence was a stronger predictor of accuracy for previously-rated items. Thus, participants in all conditions appeared attentive to the memory concepts, regardless of what task they completed prior to R2. This account also does not explain why confidence was a weaker predictor of accuracy for those in the List-Memory condition, as both the List-Memory and Explain-Memory groups were equally exposed to the memory concept. It could be that engaging in full explanations resulted in deeper learning, as testing has been found more beneficial to later learning than study alone, even for new study material (Wissman, Rawson, & Pyc, 2011). Therefore, participants, particularly those in the explanation conditions, may have benefitted from being more directly challenged on their knowledge before receiving the expert information.

Further, according to the mediator shift hypothesis (Pyc & Rawson, 2012), testing leads to improvements in retention at least in part because people realize which study strategies were inefficient and adopt new ones. Compared to those whose memory knowledge was not explicitly challenged, being forced to explain the memory items one previously rated may have uniquely improved the Explain-Memory participants' encoding strategies for memory-related material. Although overall performance did not differ given the short time delay, it is possible that the observed differences in confidence

reflect the differences in encoding quality. Future research should explore whether the observed differences in metacognitive accuracy reflect differences in attention or encoding strategies. This work should employ a more effective control task, as well as assess metacognitive performance at both immediate and delayed retrieval.

In sum, Experiment 2 provided further support that memory knowledge is susceptible to illusions of explanatory depth and that undergoing the IOED task results in the expected changes in level of claimed understanding. Further, these changes in self-reported understanding may reflect an effect of explaining in general rather than reflecting any domain-specific retrieval processes. Additionally, Experiment 2 revealed that engaging in the explanation task can result in benefits in metacognitive accuracy for those who explained the topic being tested. Although the reason why those who explained the memory concepts showed such a benefit requires further investigation, these findings do lend credence to the use of the task for improving metacognition. At present, these results do not demonstrate any differences in immediate post-task performance, calling into question the utility of the IOED task as a useful tool in contexts where time is limited or otherwise constrained. However, the unexpected results for those in the listing condition hint at the need for a better control task, and the task's effect on more applied learning remains in question. Thus, future work should test not only people's immediate knowledge of a topic, but also their ability to apply the new information to a relevant problem.

CHAPTER FOUR

Experiment Three

Overview

In Experiment 3, I explored whether the IOED paradigm can be used in a more applied setting: the courtroom. Mock jurors rated their perceived understanding of how memory works before either explaining the memory concepts or completing a filler task before reading case materials. The case materials included either strong or weak eyewitness evidence against the defendant. After receiving a set of modified jury instructions, participants evaluated the eyewitness and rendered judgments about the case in general. Findings from Experiment 3 served as an initial look into 1) the usefulness of the IOED for applying learned material and 2) an examination of the IOED task as a tool for improving prospective jurors' ability to evaluate eyewitness evidence.

Hypotheses

I had three predictions for this experiment:

1. I will confirm two major findings from Experiment 2, such that a) participants who provide explanations will experience a decrease in ratings of perceived understanding compared to those who do not and b) explaining will result in improved calibration and performance on the memory questionnaire compared to those who complete a filler task.
2. Participants in the Explain condition will differ in their ratings of eyewitness accuracy compared to those in the Control condition.

3. Although I expect to see differential responses on perceptions of guilt and verdict from those in the Explain and Control conditions, I do not have any expectations with regard to direction or statistical significance.

Method

Participants

I recruited 199 participants using MTurk for this study. All participants were paid $1.50 for their work. I once again utilized the Turk Prime toolkit for MTurk worker recruitment (Litman, Robinson, & Abberbock, 2016). Only workers with at least 1000 approved HITs and a HIT approval rate of at least 98% were allowed to take part in the study. Additionally, I recruited only participants whose location was set to Texas. The sample was predominantly White and female (63.3% White; 68.8% Female). Only one participant indicated they had not received at least a high school diploma or equivalent; 85.9% of participants noted they had completed at least some college. Participants were 38.2% Conservative-leaning, 19.6% Moderate, and 40.2% Liberal-leaning. A plurality of participants estimated their household income at between $30,000-$39,999 (17.2%), with the next-largest group indicating their income at $20,000-$29,999 (11.1%). Most participants indicated they completed the survey during the afternoon (12:00PM – 5:59PM; 39.2%). A plurality of participants estimated they had 8 hours of sleep the night before (21.1%), and a similar margin believed they needed 8 hours of sleep per night (31.2%). All participants were at least 18 years old (M_{Age} = 38.82; *SD* = 11.51; Range = 18-73) and nearly all were eligible to vote in the United States (98.0% Eligible to vote).

Materials

IOED questionnaire. I employed the same IOED questionnaire as in Experiment 2; however, some of the memory concepts participants rated were modified to fit the context of the eyewitness testimony (see Appendix C for differences between Experiments 2 and 3).

Jury instructions. As in Experiment 2, I used a modified version of the *Henderson* instructions; however, I appended these instructions with descriptions of how unconscious transference and mugshot bias can influence eyewitness memory. These additional statements were taken from Kassin and colleagues (2001). These instructions scored a 15.9 on the Flesch-Kincaid Grade Level test.

Case materials. I created a case summary for the armed robbery of a local convenience store that includes prosecution and defense statements. The prosecution evidence included the confident identification of the defendant by an eyewitness, and the strength of the prosecution's case was manipulated in a trial transcript from the cross-examination of the eyewitness. The eyewitness transcripts were based on those used by Alonzo and Lane (2010) and are available in Appendix D. In the weak version of the case against the defendant, the defense noted that the defendant frequented by the store where the eyewitness worked (unconscious transference), that the eyewitness viewed the defendant's photo in a mugbook before being asked to make an identification (mugshot bias), that the eyewitness was a different race from the defendant (cross-race bias), and that the police confirmed the eyewitness's identification (confidence malleability). In the strong version of the case, these statements by the defense are replaced with ones that

underscore the strength of the eyewitness's identification.[1] The eyewitness testimony and case materials were piloted ($N = 40$) to ensure that the case materials were ambiguous and that the case strength manipulation was successful. Participants were fairly divided in their verdicts (55.00% Guilty, 45.00% Not Guilty). Separately, those who read the strong version of the eyewitness testimony found the eyewitness more accurate than those who viewed the weak version, $t(32.12) = -2.89$, $p = .007$, Glass's $\Delta = 0.77$.

Memory questionnaire. I used the same memory questionnaire developed in Experiment 2 and indicated if they believed the statements were correct (0 = No, 1 = Yes). Participants gave prospective metamemory judgments for the concepts before they started the questionnaire.

Dependent Measures

Level of understanding. Participants once again rated their level of understanding on a 0-100% scale (0 = No understanding, 100 = Complete understanding).

Eyewitness accuracy. Participants rated the perceived accuracy of the eyewitness's identification on a 0-100% scale (0 = 0% accurate, 50 = 50% accurate, 100 = 100% accurate).

[1] In Alonzo and Lane (2010), unconscious transference, confidence malleability, and mugshot bias influenced ratings of eyewitness accuracy, particularly when the information regarding these topics indicated low eyewitness accuracy. Effect sizes for these topics ranged from $\eta_p^2 = .05$ to .35. Cross-race bias did not have a significant effect on participant ratings, although its effect was trending in the expected direction ($p < .07$).

Defendant culpability. Participants rated the likelihood that the defendant committed the crime on a 0-100% scale (0 = 0% chance defendant committed it, 50 = 50% chance defendant committed it, 100 = 100% chance defendant committed it).

Verdict. Participants indicated their verdicts dichotomously (0 = Not guilty, 1 = Guilty).

Prospective metamemory. Before beginning the memory questionnaire, all participants were asked to rate how confident they were that they could answer questions about 16 topics, to include the 8 concepts they rated for level of understanding and 8 previously unrated concepts. Participants indicated their confidence on a 0-100% scale (0 = Not at all confident, 100 = Completely confident).

Confidence in retrieved answers. After responding to each item on the memory questionnaire, participants indicated the probability that they answered the question correctly on a 50-100% scale (50 = 50% confidence, 100 = 100% confidence).

Accuracy. Participants' actual performance on the memory questionnaire was scored (0 = Incorrect, 1 = Correct).

Procedure

At the start of the study, participants were randomly assigned to one of four treatment levels before the study session in a 2 (Condition: Explain vs. Control) x 2 (Case strength: Strong vs. Weak) factorial design. Participants filled out their demographic information before beginning the study.

After reading the instructions for the task, all participants rated their perceived level of knowledge on eight memory concepts, including the four items of interest in the case summary (R1; see Figure 4.1). For the Explain condition, participants then gave explanations about four of the items; participants in the control condition were asked to complete a word fragment completion task in lieu of the explanation task. Afterward, all participants re-rated their understanding (R2).

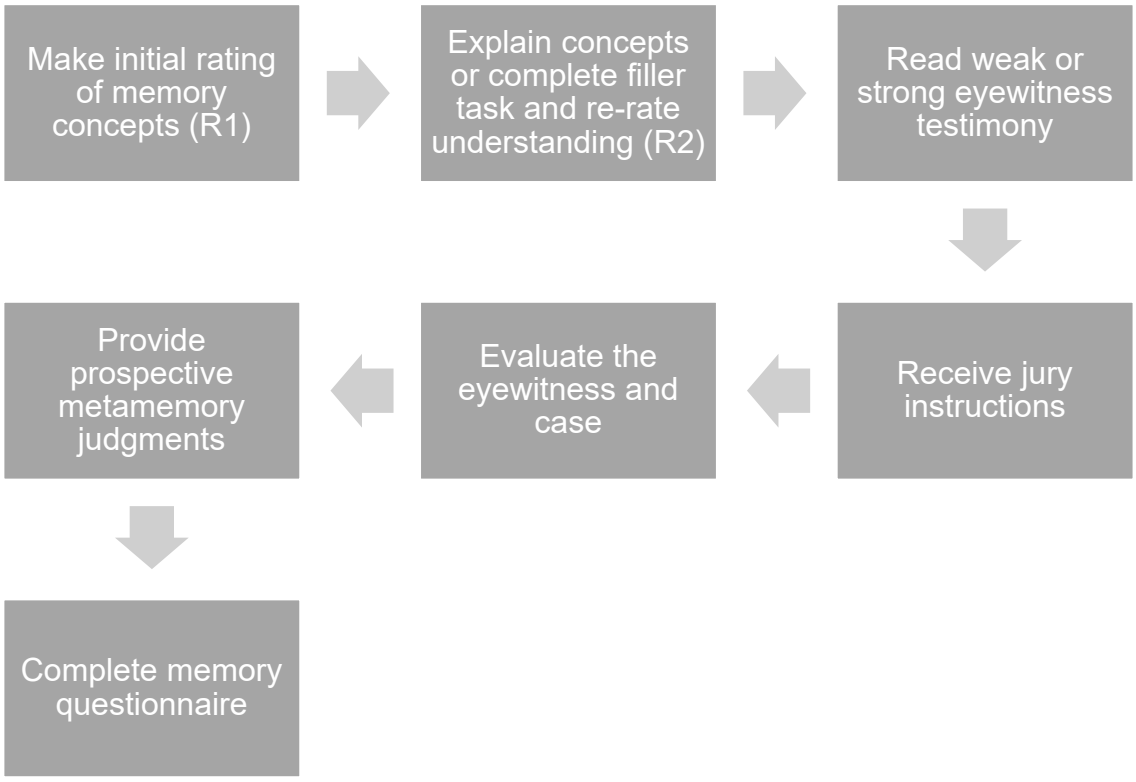

Figure 4.1. Methodology for Experiment 3.

Following their R2 ratings, participants read the case materials at their own pace, including either the strong or weak version of the eyewitness testimony against the defendant. Afterward, participants read the jury instructions. These instructions were

given after the close of evidence but before jurors made their decisions, as is common in many jurisdictions (American Bar Association, 2017). Participants then evaluated the eyewitness's accuracy, the defendant's likelihood of committing the crime, and rendered their verdicts. Finally, all participants were shown a list of 16 memory topics (half old and half new) and asked how confident they were in their ability to answer questions about the topics before completing the memory questionnaire. At the end of all experimental materials, participants were debriefed and released.

Results

Level of Understanding

To assess participants' changes in levels of perceived understanding, I analyzed the data with a linear mixed-effects model with condition (Condition: Explain or Control), Rating (1 or 2), and their interaction as fixed effects and participants and concepts as random intercepts. The Control condition and R1 served as reference groups. Participants' reported level of understanding were mean-centered by participant.

Participants in the Explain condition had higher ratings of claimed understanding for the memory concepts at baseline, $B = 3.25$, $SE = 0.77$, $\beta = 0.10$, $p < .001$, 95% CI [1.74, 4.76] (Figure 4.2). Further, participants who completed the filler task showed a decrease in ratings of understanding at the second rating, $B = -1.37$, $SE = 0.62$, $\beta = -0.04$, $p = .027$, 95% CI [-2.59, -0.16]. These main effects were qualified by a significant condition by time interaction, $B = -6.51$, $SE = 0.88$, $\beta = -0.21$, $p < .001$, 95% CI [-8.23, -4.79], such that those in the Explain condition had lower ratings than those in the Control group at the second rating. Although all participants' ratings of understanding decreased over time, participants in the Control condition ratings decreased less sharply (M_{CONTROL}-

$_{R1}$ = 47.35, *SEM* = 0.95; $M_{CONTROL-R2}$ = 45.98, *SEM* = 0.97) compared to those in the Explain condition ($M_{IOED-R1}$ = 45.65, *SEM* = 0.97; $M_{IOED-R2}$ = 37.77, *SEM* = 0.96).[2]

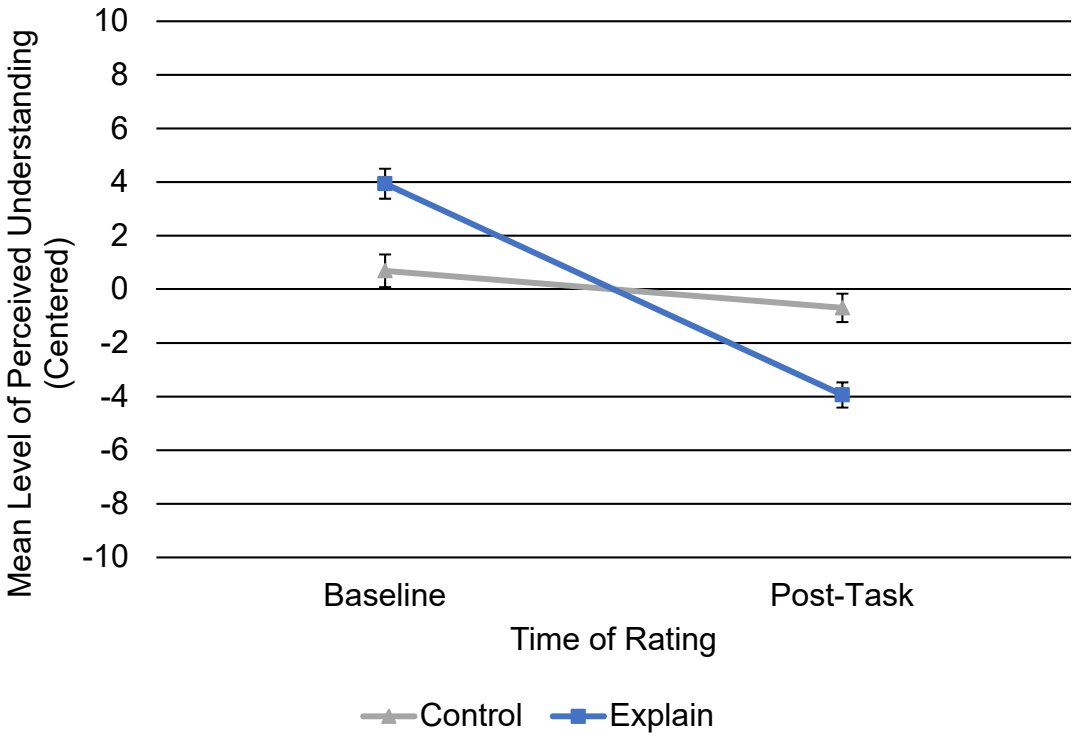

Figure 4.2. Level of perceived understanding decreased more steeply for those who explained. Note: Baseline = R1; Post-Task = R2.

Juror Decision-Making

To evaluate the effect of the manipulations on participants' reading of the case, I analyzed the data using a series of 2 (Condition: Explain or Control) x 2 (Case: Strong or Weak) factorial models. I conducted a logistic regression for verdict and separate analyses of variance models for perceived eyewitness accuracy and defendant culpability ratings.

[2] Reported means are uncentered for ease of interpretation.

Verdict. Participants were fairly split on verdict, with nearly 60% of participants finding the defendant Not Guilty (58.29%). As expected, there was a significant main effect of case strength on verdict decision, $B = 1.76$, $SE = 0.46$, $p < .001$, 95% CI [0.89, 2.69], $OR = 5.80$. Those who received the strong version of the eyewitness's testimony were more likely to find the defendant guilty. There was no significant effect of condition on verdict, $B = 0.75$, $SE = 0.46$, $p = .104$, 95% CI [-0.14, 1.69], nor was there a significant interaction of condition and case, $B = -1.01$, $SE = 0.62$, $p = .101$, 95% CI [-2.23, 0.19].

Eyewitness accuracy. There was a significant large main effect of case strength, $F(1, 195) = 41.42$, $p < .001$, partial $\eta^2 = .18$. As expected, those who received the strong version of the case found the eyewitness more accurate than those who read the weak version ($M_{STRONG} = 65.40$, $SEM = 2.41$; $M_{WEAK} = 43.40$, $SEM = 2.42$). There was also an unexpected small main effect of condition, $F(1, 195) = 8.17$, $p = .004$, partial $\eta^2 = .04$. Those who completed the filler task found the eyewitness more accurate ($M_{CONTROL} = 59.29$, $SEM = 2.41$; $M_{EXPLAIN} = 49.51$, $SEM = 2.42$). Finally, the condition by case interaction was approaching significance, $F(1, 195) = 3.37$, $p = .068$ (Figure 4.3). Those who read the strong version of the eyewitness's testimony found the eyewitness similarly accurate regardless of condition ($M_{STRONG\text{-}EXPLAIN} = 63.63$, $SEM = 3.44$; $M_{STRONG\text{-}CONTROL} = 67.16$, $SEM = 3.37$). However, those who read the weak version differed in the expected direction; participants who completed the explanation task found the eyewitness much less accurate than those who did not ($M_{WEAK\text{-}EXPLAIN} = 35.38$, $SEM = 3.40$; $M_{WEAK\text{-}CONTROL} = 51.43$, $SEM = 3.44$).

Defendant culpability. There was a significant medium effect of case on ratings of defendant culpability, $F(1, 195) = 28.71$, $p < .001$, partial $\eta^2 = .13$. In line with expectations, those who received the strong version of the case found the defendant more culpable ($M_{STRONG} = 68.23$, $SEM = 2.66$; $M_{WEAK} = 48.00$, $SEM = 2.67$). Separately, there was a small unexpected main effect of condition, $F(1, 195) = 4.51$, $p = .035$, partial $\eta^2 = .02$. Those in the control condition found the defendant more culpable ($M_{CONTROL} = 62.13$, $SEM = 2.66$; $M_{EXPLAIN} = 51.10$, $SEM = 2.67$). The interaction was not significant, $F(1, 195) = 1.27$, $p = .261$.

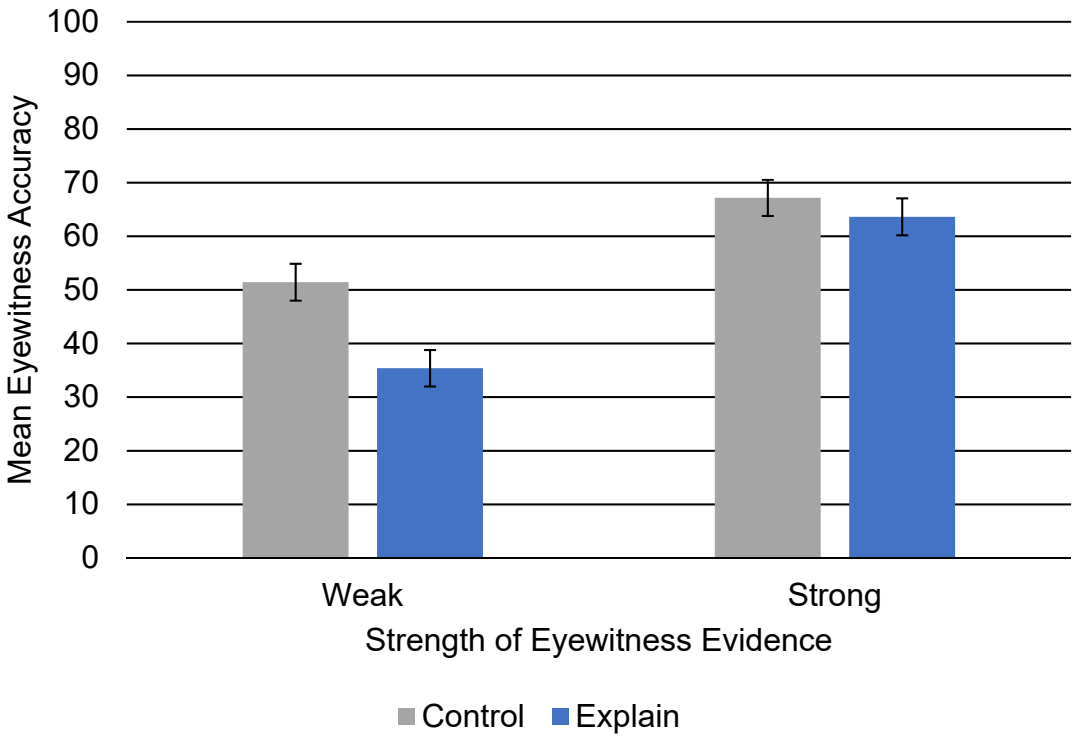

Figure 4.3. Explaining made participants more sensitive to the weak eyewitness testimony.

Performance

On average, participants answered 11.70 of the 16 questions correctly (*SEM* = 0.14; Range = 4-16). To explore any possible condition effects on performance, I conducted a linear regression with condition (Condition: Explain or Control) as a fixed effect and participants as random intercepts. The Control condition served as the reference group. Participants' overall score on the memory questionnaire did not differ between the groups, $B = -0.16$, $SE = 0.29$, $\beta = -0.04$, $p = .571$, 95% CI [-0.73, 0.40]. Participants who gave explanations performed similarly to those who completed the filler task ($M_{EXPLAIN} = 11.63$, $SEM = 0.19$; $M_{CONTROL} = 11.79$, $SEM = 0.22$).

Metacognition

To assess the utility of the instructions and the memory concepts they introduced on participants' performance, I ran mixed-effects logistic regression models with accuracy as the dependent variable, confidence (either prospective metamemory judgments or retrieval confidence) as a fixed effect, participants and items as random intercepts, and confidence judgments by participant as a random slope. When condition was included in the models, the Control condition served as the reference group. Participants' confidence in their individual judgments were mean-centered by participant.

Prospective metamemory judgments. Participants were moderately confident in their ability to answer future questions on the listed memory concepts ($M = 42.19$, $SEM = 1.78$).[3] For the eight original memory concepts, prospective metamemory judgments were

[3] Reported mean is uncentered for ease of interpretation.

positively related to initial, $r = .59$, $p < .001$, and second ratings of reported understanding, $r = .69$, $p < .001$.

To assess how the prospective metamemory judgments reflected participants' awareness of their performance, I first determined which memory topic corresponded to which questionnaire item. Due to the changes I made to the rated memory items for Experiment 3, I could not find suitable corresponding questionnaire items for two concepts from the eight memory concepts they also rated for perceived understanding: *familiarity with a person and suspect misidentification* and *factors that can inflate confidence in a memory*. Thus, I conducted the prospective metamemory judgments analyses with only the resulting 14 pairs of memory concepts and questionnaire items.

When entered alone into the basic model, centered prospective metamemory judgments were not significantly related to accuracy, $B = 0.002$, $SE = 0.004$, $p = .625$, 95% CI [-0.01, 0.01]. Adding condition (Explain or Control) and the condition by judgment interaction to the model did not change this pattern. Neither prospective metamemory judgments or condition significantly predicted questionnaire accuracy, $B = 0.003$, $SE = 0.01$, $p = .632$, 95% CI [-0.01, 0.01], and $B = -0.15$, $SE = 0.15$, $p = .347$, 95% CI [-0.45, 0.16], respectively, nor did the condition by judgment interaction, $B = -0.001$, $SE = 0.01$, $p = .873$, 95% CI [-0.02, 0.01]. Figure 4.4 shows overall prospective metamemory judgments on performance by group.

Confidence in retrieved answers. Participants were fairly confident in their answers to the memory questionnaire ($M = 82.36$, $SEM = 0.68$)[4], and uncentered confidence ratings were positively related to prospective metamemory judgments on the

[4] Reported mean is uncentered for ease of interpretation.

14 related memory concepts, $r = .28$, $p < .001$. In contrast to the prospective metamemory judgments, retrieval confidence ratings were only weakly positively related to participants' initial ($r = .08$, $p = .006$) and second ratings of perceived understanding ($r = .08$, $p = .004$) on the six memory concepts that were covered in both the ratings and the memory questionnaire.

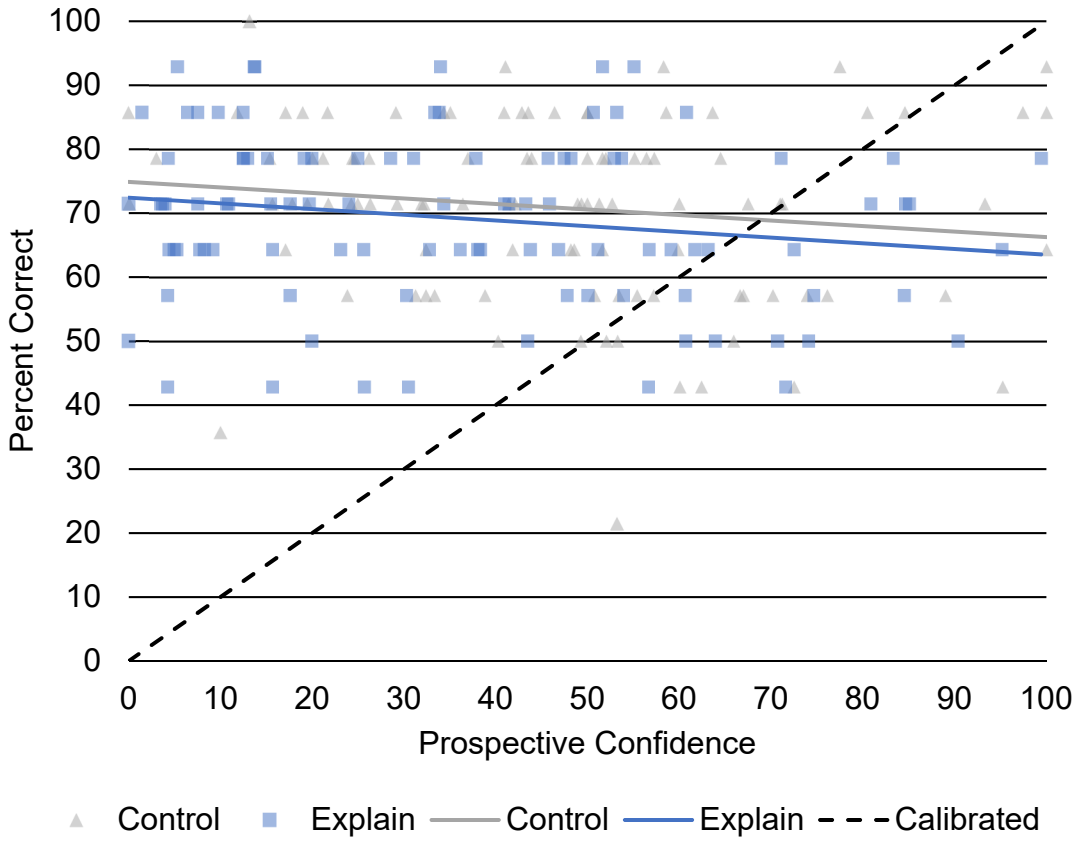

Figure 4.4. Simple calibration curves by condition for the prospective metacognitive judgments in Experiment 3. The black dashed line represents perfect calibration.

Centered confidence ratings alone positively predicted memory performance, $B = 0.02$, $SE = 0.004$, $p < .001$, 95% CI [0.01, 0.02], $OR = 1.02$. Participants were slightly more likely to get questions correct as their confidence increased. To investigate any effect of group on metacognition, I added condition (Explain or Control) and the

condition by confidence interaction to the model as fixed effects. Participants' accuracy was similar regardless of condition B = -0.13, SE = 0.16, p = .424, 95% CI [-0.45, 0.19], and participants' confidence ratings did not differentially predict accuracy by condition, B = -0.01, SE = 0.01, p = .378, 95% CI [-0.02, 0.01]. However, confidence ratings still positively predicted performance, B = 0.02, SE = 0.01, p < .001, 95% CI [0.01, 0.03], OR = 1.02. Figure 4.5 shows the overall confidence on performance by group.

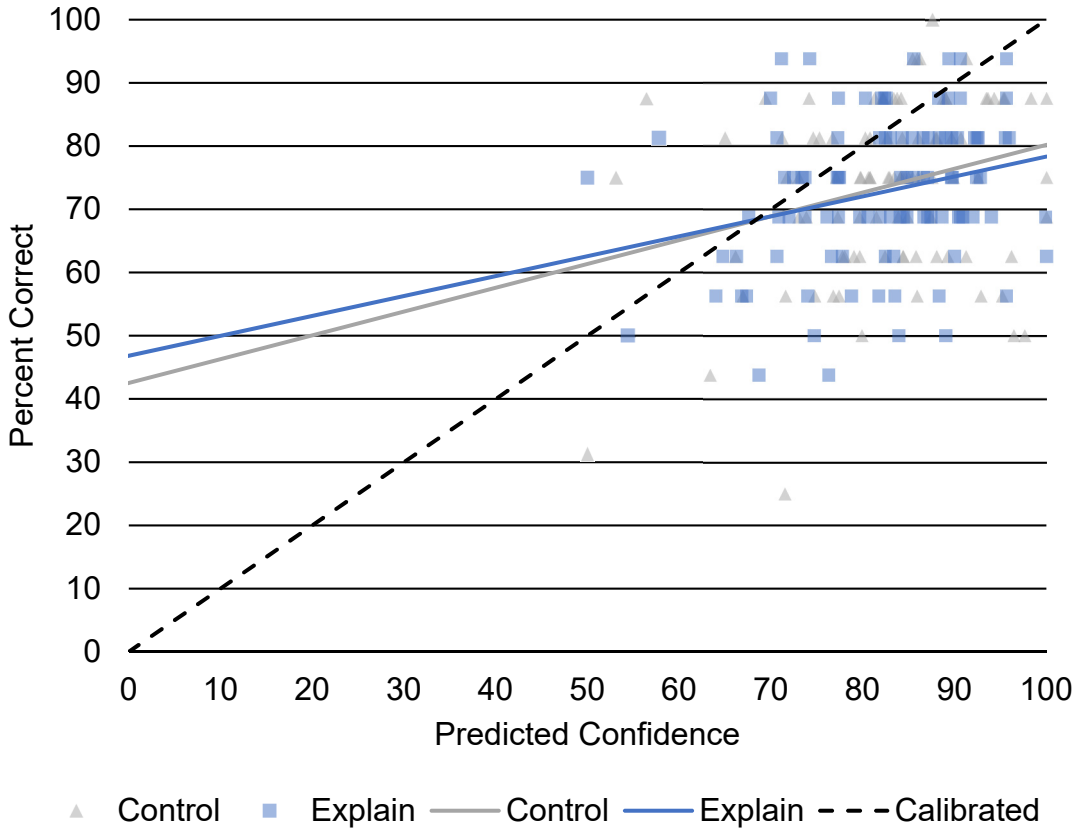

Figure 4.5. Simple calibration curves by condition for confidence in retrieved answers in Experiment 3. The black dashed line represents perfect calibration. Once again, the regression lines have been extended here for ease of comparison.

Discussion

Experiment 3 provided further support for the use of the IOED task within the memory domain. Participants who were asked to engage in the explanation task showed a steeper decrease in reported understanding for the memory concepts than those who completed the filler task, providing partial support to my first hypothesis. As in Experiment 2, participants in my Control condition still showed declines in level of perceived understanding after completing their task. This decrease in ratings may be an experimental artifact, as participants' may have felt being asked to provide their ratings a second time implied that they must need to change them (Orne, 1962). Alternatively, participants' mental representations of the scale anchors may have been skewed after repeatedly rating conceptually relevant items on the same response scale (Mochon & Frederick, 2013). However, in contrast to Experiment 2, those in the Control group showed a shallower decrease in ratings at R2, more closely replicating the findings of other IOED research that utilized a control group (see Fernbach, Rogers, et al., 2013; Zeveney and Marsh, 2016). The word fragment completion task therefore likely serves as a better control condition than asking people to list what they know or associate with memory concepts.

Importantly, in line with my second hypothesis, engaging in the explanation task appeared to influence participants' evaluations of eyewitness evidence. Participants who explained found the eyewitness less accurate in general. It is possible that participants who were challenged in their knowledge were more skeptical towards either memory specifically or information more broadly; however, it is more likely that this overall lower accuracy was due to a tendency for those in the Explain condition who read the

weak version of the eyewitness's testimony to find him less accurate than those in the Control group. Although participants in both conditions were able to discern between the strong and weak version of the eyewitness's testimony, this trending interaction hints that providing explanations may make people more sensitive to poor quality eyewitness memory.

In a similar vein, those who were challenged to explain the memory concepts were less likely to find the defendant culpable compared to those who did the filler task. Given that the prosecution's case hinged on eyewitness identification, participants in the Explain group, who were more skeptical of the eyewitness, may have been more critical of the case overall. Previous research exploring ways to educate jurors about eyewitness memory has observed that increased skepticism in eyewitness testimony can induce skepticism in the prosecution's case, especially when the information about the eyewitness is made salient (Leippe et al., 2004; Safer et al., 2016). Yet these exculpatory effects are usually accompanied by differences in verdict decisions even in the face of non-eyewitness evidence. Although participants' verdict decisions were sensitive to the quality of the eyewitness strength, those in the Control condition were no more likely to find the defendant guilty than those in the Explain condition. As participants' verdict decisions were nearly evenly split, these findings suggest that participants considered evidence other than the eyewitness's testimony when making their decisions.

Despite these findings in favor of the IOED task on self-reported understanding and the juror decision-making variables, there were no differences in memory knowledge or metacognitive ability between the two groups. Thus, the calibration and performance portions of my first hypothesis were not supported. As in Experiment 2, those who

explained performed similarly on the questionnaire to those who completed the filler task, likely due to the short delay. However, unlike Experiment 2, there were no differences in the predictive ability of participants' prospective metamemory judgments or confidence in retrieved answers by condition. Participants' prospective metamemory judgments in particular appeared to be unrelated to their ability to answer items on the questionnaire correctly. Comparatively, participants' confidence was more predictive of their actual knowledge than prospective metamemory judgments; yet engaging in the explanation task did not result in any metacognitive benefits over the filler task. It is possible that these discrepancies from Experiment 2 with regard to response confidence are due to the addition of the juror decision-making task. For example, the case description's focus on eyewitness testimony may have primed participants in both conditions to attend more to the expert information about memory that appeared near the end of the case information, and the greater delay between generating explanations and receiving feedback may have minimized any benefit previously seen for those who explain the memory concepts.

Additionally, the relationships between participants' self-reported understanding, metacognitive estimates, and their actual knowledge suggest that ratings of understanding are only weakly related to participants' metacognitive awareness, at least at immediate test. Participants' initial and secondary ratings of understanding were strongly positively correlated with prospective metamemory judgments, yet prospective metamemory judgments had no association with participants' accuracy on the questionnaire. Separately, both R1 and R2 were significantly related to confidence in answers on those items. This is in contrast to Experiment 2, where only final ratings of perceived

understanding were related to confidence judgments. However, the strength of the correlation between these ratings and confidence remained fairly small, particularly when compared to the ratings' relationship with prospective metamemory judgments.

The findings from Experiment 3 demonstrate once again that undergoing the explanation task results in changes in self-reported understanding for memory concepts. With the inclusion of the improved control task, it appears evident that memory is susceptible to an illusion of explanatory depth. Further, challenging people to explain what they claim to know renders them more likely to report lower levels of understanding than those who are not challenged in their knowledge. Undergoing the task as part of trial proceedings does not appear to lead to improvements in metacognition or actual knowledge when measures are administered relatively soon following the task; however, engaging in explanation did make people more skeptical of and perhaps more sensitive to eyewitness memory evidence. Challenging the illusion of explanatory depth for memory may therefore increase sensitivity to differences in the quality of an eyewitness's memory.

CHAPTER FIVE

General Discussion

Eyewitness misidentification is one of the most common causes of false incarceration in the United States (Innocence Project, 2019a). By educating the public on how memory works, researchers can potentially reduce prospective jurors' overreliance on memory evidence in the courtroom and make them more critical assessors of such evidence. However, interventions aimed to improve jurors' understanding of memory, such as expert testimony and pattern jury instructions about eyewitness memory, can often cause people to become skeptical of all eyewitness evidence, regardless of its quality. Although some methods, like the Interview-Identification-Eyewitness (I-I-Eyes) teaching aid (Pawlenko, Safer, Wise, & Holfeld, 2013) improve laypersons' sensitivity to the quality of an eyewitness's memory (Safer et al., 2016), they may be too time- or labor-intensive to administer within the constraints of the courtroom. Therefore, the need for a simple yet effective method of improving jurors' understanding of and sensitivity for memory remains.

I proposed a way to correct the public's misbeliefs about memory by challenging their understanding of how memory works before providing expert information. By weakening the illusions people had about the depth of their memory knowledge and forcing them to acknowledge their own ignorance, I hoped to render people more attentive to corrective, expert feedback. Over the course of three experiments, I demonstrated not only that the memory domain is susceptible to an illusion of

explanatory depth and is appropriate to use with the IOED task, but that challenging people to provide explanations results in reliable and expected changes in self-reported understanding of memory concepts. In showing that people have overconfidence in their causal understanding of memory that is challenged and re-assessed through this task, this work adds to literature by extending the application of the IOED paradigm to a new domain.

Being challenged on one's knowledge did not significantly improve performance over control conditions on a memory questionnaire administered shortly after the IOED task. However, being forced to explain as few as four concepts did improve people's evaluation of eyewitness evidence in a juror decision-making task. While preliminary, these results suggest that being forced to reassess one's knowledge about memory works can enhance how people review eyewitness memory in a relatively short period of time. Although both the Explain and Control groups found the stronger version of the eyewitness's testimony more accurate than the weak version and gave similar ratings of accuracy for the strong version, those who explained were more critical of the weaker eyewitness evidence. But why would those who provided explanations be more sensitive to the poorer quality eyewitness memory? It is improbable that being confronted with their own ignorance about memory rendered them more skeptical of all memory given their ability to discern between the strong and weak versions of the testimony. It is more likely, then, that engaging in explanations about memory causes people to process later information on the subject differently (or at least more deeply) than they would otherwise. Future research is needed to determine if explaining the memory concepts

primes people to pay more attention to the later information about memory, or if explaining in general causes people to process the later information more effectively.

These initial findings are promising for the use of the IOED paradigm in improving laypersons' evaluation of eyewitness evidence, but they present a less promising picture for the use of the IOED task to rapidly improve overall knowledge. Generating explanations has been shown previously to benefit learning and metacognition even after a short delay (e.g., Coleman, Brown, & Rivkin, 1997), but this was not observed with my administration of the IOED task. This distinction is likely due to the fact that the questionnaire was administered immediately following the IOED task in Experiment 2 and shortly after completing the juror decision-making task in Experiment 3. As previous research has found that the benefits of explanation or retrieval often only appear after a delay (see Roediger & Karpicke, 2006a), it would be worthwhile to conduct future research that explores whether any benefits on learning or metacognition from the IOED task appear only after a day or more has passed.

Separately, while I observed a benefit in metacognitive accuracy for those who explained the memory concepts in Experiment 2 over those who were instructed to list or who explained concepts in a different domain, I did not replicate this effect in Experiment 3. This discrepancy may be attributable to the methodological differences between the two experiments. For instance, in Experiment 2 all participants completed their confidence ratings shortly after their third and final rating of their level of understanding, and participants in all conditions appeared to engage in some explanatory processes, including those in the intended control condition. Participants in Experiment 3, on the other hand, had to complete a juror decision-making task between their ratings of

perceived understanding and two measures of metacognition. Further, the juror decision-making task itself likely required considerable attentional resources that may have impacted participants' ability to assess their own knowledge. Participants were required to not only read a longer version of the memory-focused jury instructions but were also asked to read several pages of case details, including the eyewitness transcript, and make judgments about those materials. Additional research is necessary to parse out whether the lack of metacognitive effects in Experiment 3 are a result of the intervening task, the increased duration between the ratings of understanding and the expert feedback, a combination of the two, or another factor entirely. At this time, any conclusions about the utility of the IOED paradigm for improving metacognition must be made with caution.

To my knowledge, this is the first investigation using the IOED paradigm in conjunction with assessments of metacognitive awareness. The ratings of understanding collected as part of the IOED task are often conceptualized as ratings of confidence (e.g., Vitriol & Marsh, 2018), but there has yet to be any formal exploration of these ratings in relation to more conventional measures of metacognition. In this series of experiments, people's self-reported changes in understanding were significantly related to people's assessments of their own knowledge. These findings suggest that people use a similar strategy to determine their level of perceived understanding as they do for their prospective metamemory judgments and confidence in retrieved responses. Notably, prospective metamemory judgments were more strongly associated with ratings of understanding than confidence in retrieved answers. Thus, ratings of understanding likely reflect people's monitoring of the acquisition or maintenance of their knowledge rather than its active retrieval (Dunlosky & Thiede, 2013).

There are several experimental limitations to this work that restrict the generalizability of the findings. First, it is important to note that I relied entirely on online samples using Amazon's MTurk for these three experiments. MTurk samples, while more diverse than more traditional university samples (Buhrmester, Kwang, & Gosling, 2011), are not necessarily nationally representative (Arditte, Çek, Shaw, & Timpano, 2016). However, the predominantly White samples seen here may not be dramatically distinct from actual jury selection samples as recent work has found jury pools tend to over-represent White Americans and under-represent minority populations (see Berner, Brown, Da Silva, Simpson, & Guindon, 2016; Joshi & Kline, 2015).

A more pressing concern is the reliance on online materials. Although previous research has found MTurk participants are less likely to respond dishonestly for general questions and are less likely to cheat when explicitly asked not to (Clifford & Jerit, 2016; Goodman, Cryder, & Cheema, 2013), many of the participants, especially in Experiment 2, submitted answers taken from elsewhere on the internet. These participants were removed from analyses when their plagiarism was apparent, but it is possible that less obvious plagiarism occurred within these three studies. Given the nature of online studies, it is also feasible that participants relied on other online resources to craft their explanations or respond to the memory questionnaires. However, overall performance was fairly stable across experiments, and I did observe differences in self-reported understanding. If participants were relying on online resources in order to appear more knowledgeable, it would not explain why their level of claimed understanding for those topics changed across ratings.

Finally, the use of only written trial materials in Experiment 3 may limit the applicability of this task for the courtroom. Despite previous work concluding that written materials result in similar outcomes to more ecologically valid methods of presentation in mock juror studies (Bornstein, 1999; Pezdek, Avila-Mora, & Sperry, 2010), there are still valid concerns about the generalizability of experimental results based entirely on written materials. However, Experiment 3 was designed to evaluate the effectiveness of the IOED task at improving people's ability to apply information, namely information about how memory works. As such, it is not necessary for findings from this specific experiment to reflect what would happen with actual jurors in more realistic simulations. The results from this experiment should instead be seen as a promising initial step, and future research should explore the use of the IOED task in more ecologically valid settings.

Together, the findings from these three experiments suggest that encouraging people to acknowledge the limits of their understanding possibly can benefit their knowledge application. By realizing that their knowledge is shallower than previously assumed, people may be more motivated to gain a deeper understanding on the topic. Thus, the IOED paradigm shows potential to be a low-cost method for improving laypersons' evaluation of eyewitness evidence.

APPENDICES

APPENDIX A

Power Analyses

Experiment 1

An observed power analysis for the 2 (Condition: Devices vs. Memory) x 4 (Rating: R1 vs. R2 vs. R3 vs. R4) mixed-model linear regression on ratings of understanding was conducted using the SIMR package in R (Version 1.0.5; Green & MacLeod, 2016). After simulating the data 200 times, the power for the effect of condition was estimated at 93.50%, 95% CI [89.14, 96.49] and the power for the effect of rating was estimated at 100.00%, 95% CI [98.17, 100.00].

Experiment 2

An observed power analysis for the 3 (Condition: Explain-Memory vs. Explain-Devices vs. List-Memory) x 3 (Rating: R1 vs. R2 vs. R3) mixed-model linear regression on ratings of understanding was conducted using the SIMR package in R. After simulating the data 200 times, the power for the effect of condition was estimated at 19.50%, 95% CI [14.25, 25.68] and the power for the effect of rating was estimated at 100.0%, 95% CI [98.17, 100.00]. For the mixed-model logistic regression on questionnaire performance, power for the effect of confidence was estimated at 100.00%, 95% CI [98.17, 100.00] and the estimated power for the effect of condition was 42.00%, 95% CI [35.01, 49.17].

Experiment 3

An observed power analysis for the 2 (Condition: Explain vs. Control) x 2 (Rating: R1 vs. R2) mixed-model linear regression on ratings of understanding was conducted using the SIMR package in R. After simulating the data 200 times, the power for the effect of condition was estimated at 99.00, 95% CI [96.43, 99.88] and the power for the effect of rating was estimated at 41.00%, 95% CI [34.11, 48.16]. For the mixed-model logistic regression on questionnaire performance, power for the effect of response confidence was estimated at 100.00%, 95% CI [98.17, 100.00] and the estimated power for the effect of condition was 10.50%, 95% CI [6.62, 15.60]. For the 2 (Condition: IOED vs. Control) x 2 (Case strength: Strong vs. Weak) ANOVAs, the estimated power for the effect of condition on eyewitness accuracy was 96.50%, 95% CI [92.92, 98.58] and 64.50% on defendant evaluations, 95% CI [57.44, 71.12]. The estimated power for the effect of case was 100.00% on eyewitness accuracy, 95% CI [98.17, 100.00], and 99.00% on defendant evaluations, 95% CI [96.43, 99.88]. For the 2 (Condition: IOED vs. Control) x 2 (Case strength: Strong vs. Weak) logistic regression on verdict decision, the effect of condition was 37.00%, 95% CI [30.30, 44.09], and the effect of case was 94.50%, 95% CI [90.37, 97.22].

APPENDIX B

Illusion of Explanatory Depth Questionnaire – Memory[1]

Concepts
How are memories created
How are memories stored
How does aging affect memory
How does amnesia work
How do memories for significant events (like 9/11) work

Expert Explanations

Below you will find the explanations of each phenomenon provided by an expert. Let's assume that the expert explanations represent higher (closer to 100%) knowledge. Please read each explanation carefully. Then, re-rate your initial level of understanding of each explained item. In other words, rate your level of understanding before you read the explanation. You should also rate your current level of understanding of the explained item. That is, how well do you feel you understand the phenomenon after you've read the explanation.

How are memories created

Please read the following explanation:

> Creating a memory begins with perception. To create a memory, you must first be paying attention. Since you cannot pay attention to everything all the time, most of what you deal with every day is simply filtered out. Only a few stimuli pass into your conscious awareness.
>
> The separate sensations that were not filtered out then travel to the part of your brain called the hippocampus. The hippocampus combines these perceptions as they were happening into one single experience. This region, as well as a part of the brain called the frontal cortex, is responsible for evaluating these different sensory inputs. They then decide if the inputs are worth remembering.

[1] Introduction example partially based on unpublished materials created by Trent Terrell and Karenna Malavanti; instructions modeled after Fernbach, Rogers, et al. (2013).

This new information then has to be encoded into a form the brain can store: electricity and chemicals. Nerve cells connect with other cells at a point called a synapse. All the action in your brain happens at these synapses, where electrical pulses carrying messages leap across gaps between cells. The electrical firing of a pulse across the gap triggers the release of chemical messengers. These chemicals are called neurotransmitters. These neurotransmitters spread out across the spaces between cells, attaching themselves to nearby cells. Each brain cell can form thousands of links like this, giving a typical brain about 100 trillion synapses. The parts of the brain cells that receive these electric impulses are called dendrites, feathery tips of brain cells that reach out to nearby brain cells.

As you learn and experience the world, changes happen at the synapses and dendrites, creating more connections in your brain. The brain organizes and reorganizes itself in response to outside input caused by your experiences, education, or training. These changes are reinforced with use, so that as you learn and practice new information, detailed circuits of knowledge and memory are built in the brain. [2]

How are memories stored

Please read the following explanation:

Once a memory is created, it must be stored. Many experts think there are three ways we store memories. The first is in the sensory stage. The next is in short-term memory. Finally, some memories are stored in long-term memory. There is no need for us to maintain everything in our brain. To protect us from the flood of information that we are exposed to on a daily basis, the different stages of human memory function as a sort of filter.

The creation of a memory begins with its perception. The registration of information during perception happens in the brief sensory stage that usually lasts only a fraction of a second. It's your sensory memory that allows a perception such as a visual pattern, a sound, or a touch to linger for a brief moment after the stimulation is over.

The sensation is then stored in short-term memory. Short-term memory has a limited capacity. It can hold about seven items for no more than 20 or 30 seconds at a time. Important information is gradually moved from short-term memory into long-term memory. The more the information is repeated or used, the more likely it is to end up eventually in long-term memory, or to be "retained."

[2] Memory. (2009). Retrieved from https://health.howstuffworks.com/human-body/systems/nervous-system/memory-info.htm

Both sensory and short-term memory are limited and decay quickly. Long-term memory can store unlimited amounts of information indefinitely.[3]

How does aging affect memory

Please read the following explanation:

As you begin to age, the connections between cells that change as you learn begin to falter. This decline begins to affect how easily you can retrieve memories. This age-dependent loss of function appears in many animals. For humans, this process begins in our 20s and tends to get worse as we reach our 50s.

Researchers have several theories about what is behind this deterioration. Most suspect that aging causes major cell loss in a tiny region in the front of the brain that leads to a drop in the production of a neurotransmitter called acetylcholine. Acetylcholine is vital to learning and memory. In addition, some parts of the brain that are essential to memory are highly vulnerable to aging. One area, called the hippocampus, loses 5 percent of its nerve cells with each passing decade -- for a total loss of 20 percent by the time you reach your 80s. In addition, the brain itself shrinks and becomes less efficient as you age. Of course, other things can happen to your brain to speed up this decline. You may have inherited some unhealthy genes, you might have been exposed to poisons, or perhaps you smoked or drank too much. All these things speed up memory decline.

As you age, some physical changes in the brain can make it more difficult to remember efficiently. The good news is that this does not mean that memory loss and dementia are inevitable. While some abilities decline with age, overall memory remains strong for most people in their 70s. Research shows that the average 70-year-old performs as well on certain cognitive tests as 20-year-olds. Many people in their 60s and 70s score better on verbal intelligence tests than younger people. Studies also have shown that many of the memory problems experienced by older people can be lessened or even reversed. Studies of nursing home residents show that patients were able to make improvements in memory when given rewards and challenges. Physical exercise and mental stimulation also can improve mental function. [4]

[3] Memory. (2009). Retrieved from https://health.howstuffworks.com/human-body/systems/nervous-system/memory-info.htm

[4] Memory. (2009). Retrieved from https://health.howstuffworks.com/human-body/systems/nervous-system/memory-info.htm

How does amnesia work

Please read the following explanation:

> The kind of amnesia most often seen in the media is called neurological amnesia. Neurological amnesia is caused by damage to the areas of our brain that create memories, the cortex and the hippocampus. These parts of the brain help convert brief sensory memories to long-term ones. The hippocampus is very important for forming new memories. The cortex stores long-term memories. When these areas are damaged, there is no pathway for information to travel. The brain cannot form new memories or retrieve some old ones. The seriousness and location of the damage determines the degree and length of the amnesia. A brief loss of oxygen flow to the brain may leave someone unable to remember only a few hours. Other damage, like a head injury, may cause long-term memory loss.
>
> Amnesia can also follow a very stressful event, like experiencing war. Doctors call this dissociative amnesia. If a stressful event is intense and lasts a long time, it can activate our adrenal glands. These glands then release cortisol and other hormones into the bloodstream. Cortisol reduces the brain's ability to change shape to form new nerve pathways during memory creation. Lengthy exposure to cortisol can harm the hippocampus. It is harder to make memories with a damaged or weakened hippocampus, which can result in amnesia at the peak of the stress. Dissociative amnesia is usually temporary, but it can also cause long-term memory loss.
>
> Amnesia patients usually only lose their memory for events and information related to themselves. Their motor skills are typically preserved. This is because motor skills are stored separately from your information about yourself. The hippocampus initially processes both types, but autobiographical memories move to the cortex. Procedural ones go to the cerebellum.[5]

How do memories for significant events like 9/11 work

Please read the following explanation:

> Detailed memories of major events, like the 9/11 terrorist attacks, are known as "flashbulb" memories. Unlike everyday memories, these memories are very vivid when you recall them.
>
> Flashbulb memories are easier to recall than ordinary memories. This is in part because they arise from highly emotional events. Like most memories, these events activate the hippocampus. (The hippocampus is a brain area important for forming new memories.) However, these events also activate the emotional memory system.

[5] Conger, C. (2008). How amnesia works. Retrieved from https://science.howstuffworks.com/life/inside-the-mind/human-brain/amnesia.htm

When an event like 9/11 happens, the brain's arousal center, the amygdala, also fires up. These emotional events are remembered more easily than neutral events.

Although flashbulb memories feel very true, there is no guarantee they are. Like all memories, flashbulb memories get worse with time. They may be more reliable over time than memories for more ordinary events, but they are not videos.

Our flashbulb memories may also be open to the power of suggestion. This is because flashbulb memories are often preserved through persistent rehearsal of the memory. Repeated recall of the memory makes it feel much more solid, like a complete story. Each time the memory is recalled, though, there is a risk that incorrect details will be added. These incorrect details then become a part of the memory.

What makes flashbulb memories special isn't how accurate or consistent they are. It's how people feel about them. Flashbulb memories are special because we are very confident in these memories. We believe that those memories are factual. Surveys find that people often cannot recall all of the facts about an event after it happens. Even when people are confident they are remembering everything correctly, they often aren't. We have a difficult time accepting that these emotional memories can fade or change.[6, 7]

---

[6] Markman, A. (2015). The consistency of flashbulb memories. *Psychology Today.* Retrieved from https://www.psychologytoday.com/blog/ulterior-motives/201506/the-consistency-flashbulb-memories

[7] Trimarchi, M. (2015). Fear: Your memory's worst enemy. Retrieved from https://health.howstuffworks.com/mental-health/human-nature/fear-your-memorys-worst-enemy.htm

APPENDIX C

Ratings of Understanding

Experiment 2

Participants in Experiment 2 provided ratings of understanding for the following memory concepts:

1. Stress and memory for an event
2. Multiple viewings of a mugshot and suspect identification
3. Alcohol intoxication and memory
4. Other-race bias and suspect identification
5. Storing of memories
6. Information learned later and memory for an event
7. Level of confidence in a memory and the accuracy of a memory
8. Officer instructions during a lineup and suspect identification

Experiment 3

Participants in Experiment 3 provided ratings of understanding for the following memory concepts:

1. Stress and memory for an event
2. Multiple viewings of a mugshot and suspect identification
3. Alcohol intoxication and memory
4. Other-race bias and suspect identification
5. Familiarity with a person and suspect misidentification

6. Information learned later and memory for an event
7. Storing of memories
8. Factors that can inflate confidence in a memory

Additionally, participants in Experiment 3 provided prospective metamemory judgments on the eight aforementioned memory concepts and eight additional memory concepts:

1. Stress and memory for an event
2. Multiple viewings of a mugshot and suspect identification
3. Alcohol intoxication and memory
4. Other-race bias and suspect identification
5. Familiarity with a person and suspect misidentification
6. Information learned later and memory for an event
7. Storing of memories
8. Factors that can inflate confidence in a memory
9. Factors that influence the stability of a memory
10. Leading questions and memory for an event
11. Memory for unexpected events
12. Trained observers and memory for an event
13. Hypnosis and remembering details
14. Amnesia and memory
15. Repressed memories
16. Lighting and memory for an event

APPENDIX D

Eyewitness Testimony

Weak Version

You are about to read direct testimony from the eyewitness, STEVEN BECKERT, who identified CHRISTOPHER JACKSON, a 27-year-old Black male, in this case. BECKERT is a 42-year-old White male who worked as an employee at the time of the robbery.

CROSS EXAMINATION (in progress):

Q. Would you please state your name for the record?

A. Steven Beckert.

Q. Did you identify a suspect in an alleged burglary that took place on December 11TH, 2006?

A. Yes, I did.

Q. Can you describe what you saw that night?

A. I was closing up the store when a guy came in and asked for a pack of cigarettes. Suddenly he had his hand in his pocket and ordered me to open the register. I didn't know if he had a weapon. He told me to get on the ground after I opened the drawer. After he took the money he ran out of the store and drove off. Once I was sure he had left, I called the police.

Q. Did you describe the perpetrator to the police?

A. Yes.

Q. How did you describe this man to the police?

A. He was a black man in his mid-20s wearing a gray hoodie.

Q. Did you identify the suspect in a lineup?

A. Yes.

Q. Is the man you identified in the courtroom?

A. Yes, the defendant.

Q. So are you absolutely certain that the person you identified is the man who robbed the store?

A. Yes I am absolutely confident. The police officer told me I picked the right person.

Q. Is it your testimony that you had never seen the defendant prior to the robbery?

A. Yes. I had never seen him before.

Q. Would you be surprised if I told you the suspect's mugshot was in the book you looked through before you made your final identification?

A. Yes.

Q. Would it also surprise you if I told you that my client frequents your store several times a week?

A. Yes. That does surprise me.

Strong Version

You are about to read direct testimony from the eyewitness, STEVEN BECKERT, who identified CHRISTOPHER JACKSON, a 27-year-old Black male, in this case. BECKERT is a 42-year-old Black male who worked as an employee at the time of the robbery.

CROSS EXAMINATION (in progress):

Q. Would you please state your name for the record?

A. Steven Beckert.

Q. Did you identify a suspect in an alleged burglary that took place on December 11TH, 2006?

A. Yes, I did.

Q. Can you describe what you saw that night?

A. I was closing up the store when a guy came in and asked for a pack of cigarettes. Suddenly he had his hand in his pocket and ordered me to open the register. I didn't know if he had a weapon. He told me to get on the ground after I opened the drawer. After he took the money he ran out of the store and drove off. Once I was sure he had left, I called the police.

Q. Did you describe the perpetrator to the police?

A. Yes.

Q. How did you describe this man to the police?

A. He was a black man in his mid-20s wearing a gray hoodie.

Q. Did you identify the suspect in a lineup?

A. Yes.

Q. Is the man you identified in the courtroom?

A. Yes, the defendant.

Q. So are you absolutely certain that the person you identified is the man who robbed the store?

A. Yes I am absolutely confident. I told the police officer when I picked him from the lineup that I was certain. I had a good look at him before he ordered me to the ground.

BIBLIOGRAPHY

Alter, A. L., Oppenheimer, D. M., & Zemla, J. C. (2010). Missing the trees for the forest: A construal level account of the illusion of explanatory depth. *Journal of Personality and Social Psychology*, *99*, 436-451. doi: 10.1037/a0020218

Alonzo, J. D., & Lane, S. M. (2010). Saying versus judging: Assessing knowledge of eyewitness memory. *Applied Cognitive Psychology*, *24*, 1245-164. doi: 10.1002/acp.1626

American Bar Association. (2017). *How Courts Work*. Retrieved from https:// www.americanbar.org/groups/public_education/resources/ law_related_education_network/how_courts_work/juryinstruct.html

Arditte, K. A., Çek, D., Shaw, A. M., & Timpano, K. R. (2016). The importance of assessing clinical phenomena in Mechanical Turk research. Psychological Assessment, 28, 684–691. doi: 10.1037/pas0000217

Benton, T. R., Ross, D. F., Bradshaw, E., Thomas, W. N., & Bradshaw, G. S. (2006). Eyewitness memory is still not common sense: Comparing jurors, judges and law enforcement to eyewitness experts. *Applied Cognitive Psychology*, *20*, 115-129.

Berman, M. K. (2015). *Eyewitness Identification Jury Instructions: Do They Enhance Evidence Evaluation?* (Doctoral dissertation). ProQuest Dissertations Publishing. (3729072)

Berner, M., Brown, D., Da Silva, J. P., Simpson, C., & Guindon, M. (2016). *A process evaluation and demographic analysis of jury pool formation in North Carolina's Judicial District 15B*. Chapel Hill, NC: The University of North Carolina at Chapel Hill.

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual Review of Psychology*, *64*, 417-444. doi: 10.1146/annurev-psych-113011-1438

Bornstein, B. H. (1999). The ecological validity of jury simulations: Is the jury still out? *Law and Human Behavior*, *23*, 75-91. doi: 10.1023/A:1022326807441

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

Bothwell, R. K., & Jalil, M. (1992). The credibility of nervous witnesses. *Journal of Social Behavior & Personality*, *7*, 581-586.

Bowers, J. M., & Bekerian, D. A. (1984). When will postevent information distort eyewitness testimony? *Journal of Applied Psychology, 69*, 466-472. doi:10.1037/0021-9010.69.3.466

Brewer, N., & Burke, A. (2002). Effects of testimonial inconsistencies and eyewitness confidence on mock-juror judgments. *Law and Human Behavior*, *26*, 353-364.

Brewer, N., & Wells, G. L. (2006). The confidence-accuracy relationship in eyewitness identification: Effects of lineup instructions, foil similarity, and target-absent base rates. *Journal of Experimental Psychology: Applied*, *12*, 11-30. doi: 10.1037/1076-898X.12.1.11

Brigham, J. C., & Bothwell, R. K. (1983). The ability of prospective jurors to estimate the accuracy of eyewitness identifications. *Law and Human Behavior*, *7*, 19–30.

Buhrmester, M. D., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*, 3–5. doi: 10.1177/1745691610393980

Butterfield, B., & Metcalfe, J. (2006). The correction of errors committed with high confidence. *Metacognition and Learning*, *1*, 69-84. doi: 10.1007/s11409-006-6894-z

Cacioppo, J. T., Petty, R. E., & Kao, C. F. (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment*, *48*, 306-307. doi: 10.1207/s15327752jpa4803_13

Camerer, C., Loewenstein, G., & Weber, M. (1989). The curse of knowledge in economic settings: An experimental analysis. *Journal of Political Economy*, *97*, 1232-1254. doi:10.2307/1831894

Clifford, S., & Jerit, J. (2016). Cheating on political knowledge questions in online surveys: An assessment of the problem and solutions. *Public Opinion Quarterly*, *80*, 858–887. doi: 10.1093/poq/nfw030

Coleman, E. B., Brown, A. L., & Rivkin, I. D. (1997). The effect of instructional explanations on learning from scientific texts. *The Journal of the Learning Sciences*, *6*, 347-365.

Commonwealth v. Walker, 92 A 3d 766 (2014)

Cooper, J., & Neuhaus, I. M. (2000). The "Hired Gun" Effect: Assessing the Effect of Pay, Frequency of Testifying, and Credentials on the Perception of Expert Testimony. *Law and Human Behavior, 24*, 149-171.

Costermans, J., Lories, G., & Ansay, C. (1992). Confidence level and feeling of knowing in question answering: The weight of inferential processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*, 142-150. doi: 10.1037/0278-7393.18.1.142

Daubert v. Merrell Dow Pharmaceuticals, Inc. 509 U.S. 579 (1993)

Deffenbacher, K. A. (1980). Eyewitness accuracy and confidence: Can we infer anything about their relationship? *Law and Human Behavior, 4*, 243-260. doi:10.1007/BF01040617

Deffenbacher, K. A., Bornstein, B. H., & Penrod, S. D. (2006). Mugshot Exposure Effects: Retroactive Interference, Mugshot Commitment, Source Confusion, and Unconscious Transference. *Law and Human Behavior, 30*, 287-307. doi:10.1007/s10979-006-9008-1

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

Dunlosky, J., & Nelson, T. O. (1994). Does the sensitivity of judgments of learning (JOLs) to the effects of various study activities depend on when the JOLs occur? *Journal of Memory and Language, 33*, 545.

Dunlosky, J., & Rawson, K. A. (2012). Overconfidence produces underachievement: Inaccurate self evaluations undermine students' learning and retention. *Learning and Instruction*, *22*, 271-280. doi: 10.1016/j.learninstruc.2011.08.003

Dunlosky, J., & Thiede, K. W. (2013). Metamemory. In D. Reisberg (Ed.), *Oxford library of psychology. The Oxford handbook of cognitive psychology* (pp. 283-298). New York, NY, US: Oxford University Press.

Dysart, J. E., Lindsay, R. C. L., MacDonald, T. K., & Wicke, C. (2002). The intoxicated witness: Effects of alcohol on identification accuracy from showups. *Journal of Applied Psychology, 87*, 170-175. doi:10.1037/0021-9010.87.1.170

Eleventh Circuit. (2016). Pattern Jury Instructions (Criminal Cases). Retrieved from http://www.ca11.uscourts.gov/pattern-jury-instructions

Fernbach, P. M., Rogers, T., Fox, C. R., & Sloman, S. A. (2013). Political extremism is supported by an illusion of understanding. *Psychological Science*, *24*, 939-946. doi: 10.1177/0956797612464058

Fernbach, P. M., Sloman, S. A., St. Louis, R., & Shube, J. N. (2013). Explanation fiends and foes: How mechanistic detail determines understanding and preference. *Journal of Consumer Research*, *39*, 1115-1131. doi: 10.1086/667782

Fifth Circuit. (2015). Pattern Jury Instructions (Criminal Cases). Retrieved from http://www.lb5.uscourts.gov/viewer/?/juryinstructions/Fifth/crim2015.pdf

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1977). Knowing with certainty: The appropriateness of extreme confidence. *Journal of Experimental Psychology*, *3*, 552-564.

Fisher, M., Goddu, M. K., & Keil, F. C. (2015). Searching for explanations: How the internet inflates estimates of internal knowledge. *Journal of Experimental Psychology: General*, *144*, 674-687. doi: 10.1037/xge000007

Fisher, M., & Keil, F. C. (2015). The curse of expertise: When more knowledge leads to miscalibrated explanatory insight. *Cognitive Science*, *40*, 1251–1269. doi: 10.1111/cogs.12280

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist*, *34*, 906-911. doi: 10.1037/0003-066X.34.10.906

Frederick, S. (2005). Cognitive reflection and decision making. *The Journal of Economic Perspectives*, *19*, 25-42.

Frye v. United States, 130 S. Ct. 307 (1923)

Garrioch, L., & Brimacombe, C. A. E. (2001). Lineup administrators' expectations: Their impact on eyewitness confidence. *Law and Human Behavior, 25*, 299-315. doi:10.1023/A:1010750028643

Goodman, J. K., Cryder, C. E., & Cheema, A. (2013), Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples. *Journal of Behavioral Decision Making*, *26*, 213-224. doi:10.1002/bdm.1753

Green, P., & MacLeod, C. J. (2016). SIMR: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, *7*, 493-498. doi: 10.1111/2041-210X.12504

Griffin, T. D., Wiley, J., & Thiede, K. W. (2008). Individual differences, rereading, and self-explanation: Concurrent processing and cue validity as constraints on metacomprehension accuracy. *Memory & Cognition, 36*, 93-103.

Harlow, C. W. (2000). *Defense Counsel in Criminal Cases*. (BJS Report No. NCJ 179023). Washington, D.C.: Bureau of Justice Statistics.

Innocence Project. (2019a). The Cases. Retrieved from https://www.innocenceproject.org/cases/

Innocence Project. (2019b). DNA Exonerations in the United States. Retrieved from https://www.innocenceproject.org/dna-exonerations-in-the-united-states/

Innocence Project. (2019c). Eyewitness Identification Reform. Retrieved from https://www.innocenceproject.org/eyewitness-identification-reform/

Insurance Institute for Highway Safety. (2017). Alcohol-impaired driving. Retrieved from http://www.iihs.org/iihs/topics/t/alcohol-impaired-driving/qanda

Jackman, T. (2016, March 14). Wrongful convictions cost California taxpayers $282 million over 24 years, study finds. *The Washington Post*. Retrieved from https://www.washingtonpost.com/news/true-crime/wp/2016/03/14/wrongful-convictions-cost-california-taxpayers-282-million-over-24-years-study-finds/

Jones, A. M., Bergold, A. N., Dillon, M. K., & Penrod, S. D. (2017). Comparing the effectiveness of *Henderson* instructions and expert testimony: Which safeguard improves jurors' evaluations of eyewitness evidence? *Journal of Experimental Criminology*, *13*, 29-52.

Joshi, A. S., & Kline, C. T. (2015, September 01). Lack of jury diversity: A national problem with individual consequences. *American Bar Association*. Retrieved from https://www.americanbar.org/groups/litigation/committees/diversity-inclusion/articles/2015/lack-of-jury-diversity-national-problem-individual-consequences/

Jurilytics. (2017, March 02). Daubert and Frye in the 50 States. Retrieved from https://jurilytics.com/50-state-overview

Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science*, *319*, 966-968. doi: 10.1126/science.1152408

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

Keil, F. C. (2003). Folkscience: Coarse interpretations of a complex reality. *TRENDS in Cognitive Sciences*, *7*, 368-373. doi: 10.1016/S1364-6613(03)00158-X

Kelemen, W. L., & Weaver, C. A., III. (1997). Enhanced memory at delays: Why do judgments of learning improve over time? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*, 1394-1409. doi:10.1037/0278-7393.23.6.1394

Kelley, C. M., & Jacoby, L. L. (1996). Adult Egocentrism: Subjective Experience versus Analytic Bases for Judgment. *Journal of Memory and Language, 35*, 157-175. doi:http://dx.doi.org/10.1006/jmla.1996.0009

Köhnken, G., & Brockmann, C. (1987). Unspecific postevent information, attribution of responsibility, and eyewitness performance. *Applied Cognitive Psychology, 1*, 197-207.

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

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

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77,* 1121–1134. doi:10.1037/0022-3514.77.6.1121

Larsen, D. P., Butler, A. C., & Roediger III, H. L. (2013). Comparative effects of test-enhanced learning and self-explanation on long-term retention. *Medical Education*, *47*, 674-682. doi: 10.1111/medu.12141

Leippe, M. R., Eisenstadt, D., & Rauch, S. M. (2009). Cueing Confidence in Eyewitness Identifications: Influence of Biased Lineup Instructions and Pre-Identification Memory Feedback under Varying Lineup Conditions. *Law and Human Behavior, 33*, 194-212.

Leippe, M. R., Eisenstadt, D., Rauch, S. M., & Seib, H. M. (2004). Timing of Eyewitness Expert Testimony, Jurors' Need for Cognition, and Case Strength as Determinants of Trial Verdicts. *Journal of Applied Psychology, 89*, 524-541. doi:10.1037/0021-9010.89.3.524

Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? *Organizational Behavior and Human Performance, 20*, 159-183. doi: 10.1016/0030-5073(77)90001-0

Lichtenstein, S., & Fischhoff, B. (1978). *Training for Calibration* (Report No. TR-78-A32). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Lindsay, R. C. L., Nosworthy, G. J., Martin, R., & Martynuck, C. (1994). Using mug shots to find suspects. *Journal of Applied Psychology, 79*, 121-130. doi:10.1037/0021-9010.79.1.121

Litman, L., Robinson, J., & Abberbock, T. (2016). TurkPrime. com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods*, *49*, 433-442. doi: 10.3758/s13428-016-0727-z

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

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

Loftus, E. F., Miller, D. G., & Burns, H. J. (1978). Semantic integration of verbal information into a visual memory. *Journal of Experimental Psychology: Human Learning and Memory, 4*, 19-31. doi:10.1037/0278-7393.4.1.19

Loftus, E. F., & Palmer, J. C. (1974). Reconstruction of automobile destruction: An example of the interaction between language and memory. *Journal of Verbal Learning and Verbal Behavior, 13*, 585-589.

Luus, C. A. E., & Wells, G. L. (1994). The malleability of eyewitness confidence: Co-witness and perseverance effects. *Journal of Applied Psychology, 79*, 714-723. doi:10.1037/0021-9010.79.5.714

Malavanti, K. F. (2014). *Comparing the efficacy of expert testimony and detailed juror instructions under high and low cognitive load* (Unpublished doctoral dissertation). Baylor University, Waco, TX.

Malavanti, K. F., Terrell, J. T., Dasse, M. N., & Weaver, C. A., III. (2014). The "Curse of Knowledge" in estimating jurors' understanding of memory: Attorneys know more about memory than the general population. *Applied Psychology in Criminal Justice*, *10*, 99-105

McAuliff, B. D., & Kovera, M. B. (2008). Juror need for cognition and sensitivity to methodological flaws in expert evidence. *Journal of Applied Social Psychology, 38*, 385-408. doi:10.1111/j.1559-1816.2007.00310.x

McDaniel, M. A., Roediger, H. L., & McDermott, K. B. (2007). Generalizing test-enhanced learning form the laboratory to the classroom. *Psychonomic Bulletin & Review*, *14*, 200-206. doi: 10.3758/BF03194052

Meissner, C. A., & Brigham, J. C. (2001). Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychology, Public Policy, and Law*, *7*, 3-35.

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.

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

Metcalfe, J. (1998). Cognitive optimism: Self-deception or memory-based processing heuristics? *Personality and Social Psychology Review*, *2*, 100-110.

Metcalfe, J., Schwartz, B. L., & Joaquim, S. G. (1993). The cue-familiarity heuristic in metacognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*, 851-861. doi:10.1037/0278-7393.19.4.851

Mochon, D., & Frederick, S. (2013). Anchoring in sequential judgments. *Organizational Behavior and Human Decision Processes*, *122*, 69-79. doi: 10.1016/j.obhdp.2013.04.002

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review, 115,* 502–517. doi:10.1037/0033-295X.115.2.502

Mulligan, N. W., & Peterson, D. J. (2015). The negative testing and negative generation effects are eliminated by delay. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*, 1014-1025. doi: 10.1037/xlm0000070

National Institute of Justice. (2003). *Eyewitness evidence: A trainer's manual for law enforcement.* (NCJ 188678). Washington, D.C.: U.S. Department of Justice, Office of Justice Programs.

National Research Council. (2014). *Identifying the culprit: Assessing eyewitness identification*. Retrieved from http://www.nap.edu/catalog.php?record_id=18891.

Nelson, T. O., & Dunlosky, J. (1991). When People's Judgments of Learning (JOLs) Are Extremely Accurate at Predicting Subsequent Recall: The "Delayed-JOL Effect". *Psychological Science, 2*, 267-270.

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. *The Psychology of Learning and Motivation*, *26*, 125-173.

Orne, M. T. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist, 17*, 776-783. doi: 10.1037/h0043424

O'Sullivan, J. T., & Howe, M. L. (1995). Metamemory and memory construction. *Consciousness and Cognition*, *4*, 104-110.

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

Papailiou, A. P., Yokum, D. V., & Robertson, C. T. (2015). The novel New Jersey eyewitness instruction induces skepticism but not sensitivity. *PLOSOne*, *10*, e0142695. doi: 10.1371/journal.pone.0142695

Pashler, H., Cepeda, N. J., Wixted, J. T., & Rohrer, D. (2005). When does feedback facilitate learning of words? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 3-8. doi: 10.1037/0278-7393.31.1.3

Pawlenko, N. B., Safer, M. A., Wise, R. A., & Holfeld, B. (2012). A teaching aid for improving jurors' assessments of eyewitness accuracy. *Applied Cognitive Psychology*, *27*, 190-197. doi: 10.1002/acp.2895

Pezdek, K., Avila-Mora, E., & Sperry, K. (2010). Does trial presentation medium matter in jury simulation research? Evaluating the effectiveness of eyewitness expert testimony. *Applied Cognitive Psychology*, *24*, 673-690. doi: 10.1002/acp.1578

Pennington, N., & Hastie, R. (1992). Explaining the evidence: Tests of the Story Model for juror decision making. *J Pers Soc Psychol, 62*, 189-206. doi:10.1037/0022-3514.62.2.189

Perfect, T. J., Watson, E. L., & Wagstaff, G. F. (1993). Accuracy of confidence ratings associated with general knowledge and eyewitness memory. *Journal of Applied Psychology*, *78*, 144-147. doi: 10.1037/0021-9010.78.1.144

Pyc, M. A., & Rawson, K. A. (2012). Why is test-restudy practice beneficial for memory? An evaluation of the mediator shift hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*, 737-746. doi: 10.1037/a0026166

Read, J. D., Yuille, J. C., & Tollestrup, P. (1992). Recollections of a robbery: Effects of arousal and alcohol upon recall and person identification. *Law and Human Behavior, 16*, 425-446. doi:10.1007/BF02352268

Roeder, S. S., & Nelson, L. D. (2015). *Folk theories are corrupted by cross-domain explanations.* Retrieved from SSRN: https://ssrn.com/abstract=2622301

Roediger, H. L., & Karpicke, J. D. (2006a). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, *1*, 181-210. doi: 10.1111/j.1745-6916.2006.00012.x

Roediger, H. L., & Karpicke, J. D. (2006b). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, *17*, 249-255. doi: 10.1111/j.1467-9280.2006.01693.x

Roediger, H. L., & Pyc, M. A. (2012). Inexpensive techniques to improve education: Applying cognitive psychology to enhance educational practice. *Journal of Applied Research in Memory and Cognition*, *1*, 242-248. doi: 10.1016/j.jarmac.2012.09.002

Ross, D. R., Ceci, S. J., Dunning, D., & Toglia, M. P. (1994). Unconscious transference and mistaken identity: When a witness misidentifies a familiar but innocent person. *Journal of Applied Psychology, 79*, 918-930. doi:10.1037/0021-9010.79.6.918

Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science*, *26*, 521-562. doi: 10.1207/s15516709cog2605_1

Safer, M. A., Murphy, R. P., Wise, R. A., Bussey, L., Millett, C., & Holfeld, B. (2016). Educating jurors about eyewitness testimony in criminal cases with circumstantial and forensic evidence. *International Journal of Law and Psychiatry*, *47*, 86-92. doi: 10.1016/j.ijlp.2016.02.041

Schmechel, R. P., O'Toole, T. P., Easterly, C., & Loftus, E. F. (2006). "Beyond the ken?" Testing jurors' understanding of eyewitness reliability evidence. *Jurimetrics*, *46*, 177-214.

Shaw, J. S., Garven, S., & Wood, J. M. (1997). Co-witness information can have immediate effects on eyewitness memory reports. *Law and Human Behavior, 21*, 503-523.

Silver, J., & Carbonell, L. (2016, June 24). Wrongful Convictions Have Cost Texans More Than $93 Million. *The Texas Tribune*. Retrieved from https://www.texastribune.org/2016/06/24/wrongful-convictions-cost-texans-over-93-million/

Simmonsen, D. (2011). Teach your jurors well: Using jury instructions to educate jurors about factors affecting the accuracy of eyewitness testimony. *Maryland Law Review*, 70, 1044-1092.

Simons, D. J., & Chabris, C. F. (2011). What people believe about how memory works: A representative survey of the U.S. population. *PLoS ONE, 6*, e27757. doi: 10.1371/journal.pone.0022757

Simons, D. J., & Chabris, C. F. (2012). Common (mis)beliefs about memory: A replication and comparison of telephone and Mechanical Turk survey methods. *PLoS ONE*, *7*, e51876. doi: 10.1371/journal.pone.0051876

Simons, D. J., & Levin, D. T. (1998). Failure to detect changes to people during a real-world interaction. *Psychonomic Bulletin & Review, 5*, 644-649. doi:10.3758/BF03208840

Sirota, M., & Juanchich, M. (2018). Effect of response format on cognitive reflection: Validating a two- and four-option multiple choice question version of the Cognitive Reflection Test. *Behavior Research Methods*, *50*, 2511-2522. doi: 10.3758/s13428-018-1029-4

Son, L. K., & Kornell, N. (2010). The virtues of ignorance. *Behavioural Processes*, *83*, 207-212. doi: 10.1016/j.beproc.2009.12.005

Sparrow, B., Liu, J., & Wegner, D. M. (2011). Google effects on memory: Cognitive consequences of having information at our fingertips. *Science*, *333*, 776-778. doi: 10.1126/science.1207745

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

State v. Guilbert, 49 A.3d 705 (2012)

State of New Jersey v. Larry R. Henderson, 397 N.J. Super. 398 (2011)

Steblay, N. M. (1997). Social influence in eyewitness recall: A meta-analytic review of lineup instruction effects. *Law and Human Behavior*, *21*, 283-297.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2013). Assessing miserly information processing: An expansion of the Cognitive Reflection Test. *Thinking & Reasoning*, *20*, 147-168. doi: 10.1080/13546783.2013.844729

Truman, J. L., & Rand, M. R. (2010). *Criminal Victimization, 2009*. (BJS Report No. NCJ 231327). Washington, D.C.: Bureau of Justice Statistics.

United States Courts. (2017). *Juror Qualifications*. Retrieved from http://www.uscourts.gov/services-forms/jury-service/juror-qualifications

Van den Broek, G. S. E., Segers, E., Takashima, A., & Verhoeven, L. (2014). Do testing effects change over time? Insights from immediate and delayed retrieval speed. *Memory*, *22*, 803-812. doi: 10.1080/09658211.2013.831455

Van Knippenberg, A., Dijksterhuis, A., & Vermeulen, D. (1999). Judgement and memory of a criminal act: The effects of stereotypes and cognitive load. *European Journal of Social Psychology*, *29*, 191-201.

Van Zandt, T. (2000). ROC curves and confidence judgments in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 582-600. doi: 10.1037/0278-7393.26.3.582

Vitriol, J. A., & Marsh, J. K. (2018). The illusion of explanatory beliefs and endorsement of conspiracy beliefs. *European Journal of Social Psychology*, *48*, 955-969. doi: 10.1002/ejsp.2504

Walsh, J. T. (1998). The evolving standards of admissibility of scientific evidence. *General Practice, Solo & Small Firm*, *2*, n.p.

Weaver, C. A., & Kelemen, W. L. (1997). Judgments of Learning at Delays: Shifts in Response Patterns or Increased Metamemory Accuracy? *Psychological Science, 8*, 318-321.

Wegner, D. M. (1987). Transactive memory: A contemporary analysis of the group mind. In B. Mullen and G. R. Goethals (Eds.), *Theories of Group Behavior* (pp.185-208). New York, NY: Springer.

Wells, G. L., & Bradfield, A. L. (1998). "Good, you identified the suspect": Feedback to eyewitnesses distorts their reports of the witnessing experience. *Journal of Applied Psychology, 83*, 360-376. doi:10.1037/0021-9010.83.3.360

Wells, G. L., Memon, A., & Penrod, S. D. (2006). Eyewitness evidence: Improving its probative value. *Psychological Science in the Public Interest*, *7*, 45-75. doi: 10.1111/j.1529-1006.2006.00027.x

Wissman, K. T., Rawson, K. A., & Pyc, M. A. (2011). The interim test effect: Testing prior material can facilitate the learning of new material. *Psychological Bulletin & Review*, *18*, 1140-1147. doi: 10.3758/s13423-011-0140-7

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

Zeveney, A. S., & Marsh, J. K. (2016). The illusion of explanatory depth in a misunderstood field: The IOED in mental disorders. In A. Pagafragou, D. Grodner, D. Mirman, & J. C. Trueswell. (Eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 1020-1025). Austin, TX: Cognitive Science Society.