***** ACCEPTED FOR PUBLICATION IN 'JOURNAL OF APPLIED RESEARCH IN MEMORY AND COGNITION' ON NOVEMBER 24 2021 *****

Long Retention Intervals Impair the Confidence-Accuracy Relationship for Eyewitness Recall

Emily R. Spearing and Kimberley A. Wade

Department of Psychology, University of Warwick, UK

Email Addresses

e.spearing@warwick.ac.uk, k.a.wade@warwick.ac.uk

Correspondence concerning this article should be addressed to Kimberley A. Wade, email: k.a.wade@warwick.ac.uk

Abstract

A growing body of research suggests that confidence judgements can provide a useful indicator of memory accuracy under some conditions. One factor known to affect eyewitness accuracy, yet rarely examined in the confidence-accuracy literature, is retention interval. Using calibration analyses, we investigated how retention interval affects the confidence-accuracy relationship for eyewitness recall. In total, 611 adults watched a mock crime video and completed a cued-recall test either immediately, after 1 week, or after 1 month. Long (1 month) delays led to lower memory accuracy, lower confidence judgements, and impaired the confidence-accuracy relationship compared to shorter (immediate and 1 week) delays. Long-delay participants who reported very high levels of confidence tended to be over-confident in the accuracy of their memories compared to other participants. Self-rated memory ability, however, did not predict eyewitness confidence or the confidence-accuracy relationship. We discuss the findings in relation to cue-utilization theory and a retrieval-fluency account.

Keywords: eyewitness memory, confidence, accuracy, metacognition, calibration

General Audience Summary

Legal decision makers (e.g., judges, lawyers, jurors) have a tendency to believe that highly confident eyewitnesses are highly accurate. Many studies have explored the conditions under which a witness's confidence judgement is a reliable indicator of their memory accuracy. One factor known to reduce the accuracy of people's memory reports is the length of time between when an event is witnessed and when the witness is asked to provide a statement—known as the *retention interval*. Yet few studies have examined how different retention intervals, particularly longer delays of 1 month or more, might affect the usefulness of witnesses' confidence judgements for assessing the accuracy of their testimony. We examined how three different retention intervals affect the accuracy of witnesses' memory reports and their confidence in these reports. In total, 611 adults watched a mock crime video and then completed a memory test for the event either immediately, after 1 week, or after 1 month. Long (1 month) delays reduced the accuracy of participants' memory reports and their confidence in those reports. Furthermore, participants who completed the memory test after 1 month tended to be over-confident in their reports: Put simply, when long-delay participants reported being 80-100% confident, their memory reports were only accurate, on average, 60-70% of the time. Participants who completed the memory test immediately or after a short delay did not show this pattern. We also examined the relationship between people's beliefs about their own memory ability, and how accurate and confident they were. How people felt about their memory ability did not significantly influence their confidence judgements or the usefulness of confidence judgements for predicting eyewitness accuracy. Our findings highlight the detrimental effects that long delays can have on eyewitness memory and the importance of collecting witness statements quickly.

Introduction

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

The few studies that have investigated the influence of retention interval on the confidence-accuracy relationship suggest that longer retention intervals may reduce the

correlation between a witness's confidence and their memory accuracy compared to short delays (Horry, Colton & Williamson, 2014; Odinot & Wolters, 2006; Odinot, Wolters & van Giezen, 2013). In one study, participants viewed a videotape depicting a car accident, and following a delay, answered open-ended questions about it. A 5-week delay led to lower confidence in correct responses and a weaker confidence-accuracy correlation ($\gamma = .49$) than a 1- week ($\gamma = .63$) or 3-week delay ($\gamma = .58$, Odinot & Wolters, 2006). Similarly, when participants answered questions about a mock crime after a 1-week delay, they were less able to discriminate between correct and incorrect responses than those who answered questions immediately (Horry et al., 2014). Although these studies provided valuable insight into the impact of different delays on witnesses' confidence judgements, confidence-accuracy calibration – where accuracy is calculated for each level of confidence – was not examined (the sample sizes in Odinot & Walters, $N = 67$, and in Horry et al., $N = 80$, preclude calibration analyses). Thus, it remains unclear whether longer delay periods impair the confidence-accuracy relationship by inducing participants to become more under-confident or over-confident in their memories (Juslin, Olsson & Winman, 1996).

To guide our understanding of why and how longer retention intervals might impair confidence-accuracy calibration, we can draw on Koriat's (1997) cue-utilization theory. According to this model, people use multiple cues to make confidence judgements, including information derived from the process of remembering (i.e., experience-based cues) and their personal metamemorial beliefs (i.e., theory-based cues). One example of an experienced-based cue is the ease with which a target event is recalled. Research shows people tend to assign higher confidence ratings to information that is recalled with relative ease compared to information that requires more cognitive effort to bring to mind (Lindholm, Jonsson & Liuzza, 2018). Such experienced-based cues are typically useful indicators of memory accuracy, but sometimes these cues can lead to errors (Horry et al., 2014; Shaw & McClure,

1996). Theory-based cues are, by contrast, any cue related to a person's beliefs about how their memory works. For instance, factors relating to the witnessing situation, such as recalling that the crime was observed in broad daylight, or factors relating to the testing situation such as biased feedback from an interviewer (e.g., "you're spot on"), can lead witnesses to provide more, or less, conservative confidence judgements (Semmler, Brewer & Wells, 2004). The usefulness of theory-based cues hinges on whether a witnesses' metamemorial beliefs are accurate: Recent research suggests that participants can adjust their confidence appropriately for some 'common-sense' factors, such as the influence of lighting on memory (Spearing & Wade, in press), but not for other, poorly understood factors, such as the influence of marijuana intoxication on memory (Cormia, Shapland, Rasheed & Pezdek, 2020; Desmarais & Read, 2011; Ost, Easton, Hope, French, & Wright, 2017; Pezdek, Abed, & Reisberg, 2020).

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

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

There is, however, at least one factor that could mediate the impact of retention intervals on the confidence-accuracy relationship: people's beliefs about the reliability of their own memory (Leippe et al., 2009). People who believe that they have a highly reliable memory may feel more confident in their ability to accurately remember details following a long delay period than people who believe that they have an unreliable memory. After a short delay, however, even people with low self-rated memory ability may feel relatively confident in their ability to remember details accurately. Thus, we might expect self-rated memory ability to have a stronger influence on people's confidence judgments when the delay period is long, compared to when the delay period is relatively short. Previous research supports this prediction, showing that how people feel about their memory ability affects their confidence in their memory, and this effect is larger when their memory is relatively weak. For example, receiving feedback about one's memory ability has a larger influence on confidence when internal cues are weak than when internal cues are strong (Charman, Carlucci, Vallano & Gregory, 2010; Iida et al., 2020). This account may explain why one previous study found that perceived memory ability did not predict accuracy or confidence when participants' memories were tested after a short 5-minute delay (Saraiva et al., 2020). We are not aware of

any studies that have examined the influence of perceived memory ability after a more forensically relevant delay period.

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

Method

Participants & Design

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

In total, we recruited 778 adults through Amazon's Mechanical Turk (MTurk) using the CloudResearch platform (Litman, Robinson & Abberbock, 2017). Participants received

$2.50 upon completing the experiment. We excluded those who failed an attention check question ($N = 37$), reported skipping or watching the video more than once ($N = 57$), reported being distracted during the experiment ($N = 36$), experienced technical difficulties (N = 6), or did not complete Part 2 within 3 days of receiving the invitation link ($N = 25$). A further 6 people (3 short delay, 3 long delay) were excluded because they failed to answer any of the memory test questions correctly. The final sample consisted of 611 adults (319 women, 286 men, 6 undisclosed, M = 42.79 years, SD = 12.69, range = 18-83). There were 202 people in each of the immediate and short conditions, and 207 in the long condition, producing 9,165 observations in total. The Department of Psychology Research Ethics Committee at the University of Warwick approved this research. This study was pre-registered; the numeric data and corresponding R code are available on the Open Science Framework: https://osf.io/sgb4z/.

Materials

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

"Forget to pay a bill on time") on a 5-point scale ranging from "All the time" to "Never", with higher scores indicating better self-rated memory ability.

Narcissistic Personality Inventory (*NPI-16;* Ames, Rose & Anderson, 2006). The NPI-16 consists of 16 pairs of statements, each including a narcissistic statement (e.g., "I like to be the centre of attention") and a non-narcissistic statement (e.g., "It makes me uncomfortable to be the centre of attention"; $\alpha = .72$). Respondents indicate the statement that describes them best, with higher scores indicating higher levels of narcissism. We included the NPI-16 because we were initially interested in whether individual differences in personality would affect the amount of under- or over-confidence in witness's memory reports. As the analyses including the NPI-16 are not relevant to our main research question, we report these results in the supplementary materials.

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

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

Procedure

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

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

Data Coding

Participants' answers on the memory test were coded by the first author (ES) as either correct, incorrect or "don't know". Responses were coded as correct if they correctly

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

Results

Preliminary Analyses

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

Second, we examined the means for accuracy and confidence across delay conditions. "Don't know" responses were excluded (n = 598, 6.5% of responses) and accuracy was calculated as the number of correct responses divided by the number of correct and incorrect responses (total n = 8,567). The non-overlapping 95% confidence intervals show that short-delay participants (M = 48.00, CI = 45.90, 50.10*)* were significantly less accurate than immediate-delay participants (M = 65.83, CI = 63.53, 68.12) and significantly more accurate

than long-delay participants (*M* = 42.86, *CI* = 40.79, 44.95). Thus, even though short- and long-delay participants provided more 'don't know' responses, the accuracy of their reports was impaired compared to immediate-delay participants. The same pattern was observed for confidence. Short-delay participants (*M* = 38.82, *CI* = 36.11, 41.45) were significantly less confident than immediate-delay participants (*M* = 69.21, *CI* = 67.10, 71.32, but significantly more confident than long-delay participants (*M* = 32.84*, CI* = 30.12, 35.58).

Finally, we checked that calibration was similar for the car theft and mugging events. We created calibration curves by plotting accuracy against 4 levels of confidence (0 – 20, 30 – 40, 50 – 70, 80 – 100). Figure 1 shows that accuracy at 50-70% and 80-100% confidence was higher for the car theft scenario than the mugging scenario, but the overall pattern of calibration was similar for the two events. The count data for all calibration curves are provided in Table S2 in the supplementary materials.

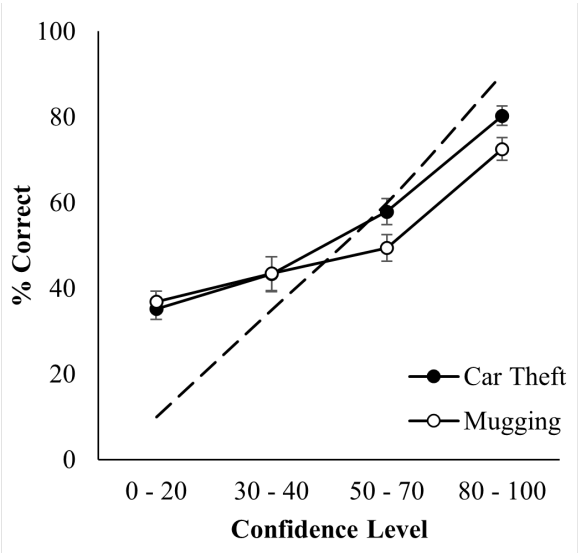

Figure 1. Calibration for each event. The dashed line represents perfect calibration. Error bars denote the 95% CI around the mean.

Main Analyses

Turning to our key research question: How does retention interval influence the confidence-accuracy relationship for eyewitness recall? To answer this question, we collapsed the data over the two events and then plotted calibration curves for each delay condition. The calibration curves were created by plotting accuracy (i.e., the proportion correct) against 4 levels of confidence (0 – 20, 30 – 40, 50 – 70, 80 – 100) so that each bin contained at least 200 observations. Figure 2 shows that the pattern of calibration was generally similar across delay conditions, with some under-confidence at the low levels of confidence and over-confidence at the highest levels of confidence. The only significant difference was at the highest level of confidence. Even though long-delay participants gave relatively few responses with 80-100% confidence (N = 317), they were more overconfident in these responses than were short-delay (N = 1,534) and immediate-delay participants (N = 491). Summary statistics and calibration statistics are reported in the supplementary materials.

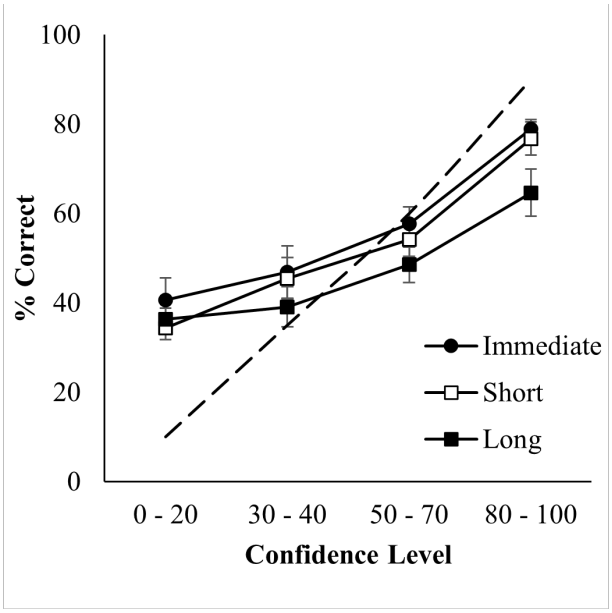

Figure 2. Calibration for each delay condition. The dashed line represents perfect calibration. Error bars denote the 95% CI around the mean.

Do people's beliefs about the reliability of their own memory influence their confidence judgements, and if so, does this depend on the length of the retention interval? To answer these questions, we conducted two multiple regressions on mean confidence and over/under-confidence including Delay, MMQ-Satisfaction, MMQ-Ability and MMQ-Strategy as predictors (Table 1). We ran the regression models in R 4.0.5 and controlled the false discovery rate by applying the Benjamini-Hochberg correction to all p-values (Benjamini & Hochberg, 1995). The models were significant for confidence, R^2 =.40, adjusted $p < .001$, and over/under-confidence, R^2 =.09, adjusted $p < .001$. Both short and long delays were associated with lower confidence judgments and greater under-confidence compared to when the memory test was taken immediately. Importantly, this underconfidence was driven by the relatively high number of responses at low levels of confidence in the two delayed conditions.

We suggested that the relationship between retrieval fluency and memory accuracy may break down over time which should lead long-delay participants to show a weaker confidence-accuracy relationship (if their confidence decisions are based, to some extent, on retrieval fluency). Consistent with this mechanism, long-delay participants showed overconfidence at high levels of confidence, but they also gave fewer high confidence responses than did immediate-delay participants, which meant that their overall (mean) decrease in confidence was larger than was justified by their lower (mean) accuracy. None of the metamemory scales predicted confidence or over/under-confidence, nor did they significantly interact with delay (adjusted $p > .05$). Additional analyses, including narcissism as a predictor of confidence, are reported in the supplementary materials.

Exploratory Analyses

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

Table 1

Results of multiple regression models with delay.

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

Note. P-values adjusted with the Benjamini-Hochberg procedure.

Discussion

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

Given that laypeople are generally aware of the negative effects of longer delays on memory performance, why did long-delay participants show over-confidence? Guided by Koriat's (1997) cue-utilization theory, we suggested that the relationship between retrieval fluency and accuracy may break down over time, and that this may lead to poorer calibration when witnesses are questioned after a relatively long delay if people continue to rely on retrieval fluency to guide their confidence judgements. Consistent with this idea, our exploratory analyses revealed that response times were significantly related to accuracy in the immediate-delayed condition but not in the short- or long-delay conditions. In contrast, response times were significantly related to confidence regardless of delay. In other words, people gave higher confidence judgements to fast responses than to relatively slow responses, even when response times did not provide a good indicator of accuracy. These findings may explain why long-delay participants showed overconfidence. Specifically, long-delay participants may have interpreted relatively high retrieval fluency as a sign that responses were likely to be correct and, as such, gave higher confidence judgements than were justified by the accuracy of those responses. This explanation is consistent with previous research showing that the confidence-accuracy relationship breaks down when retrieval fluency

provides a misleading cue to accuracy. For example, when people are exposed to post-event misinformation, this information tends to come to mind more quickly and, as such, is sometimes reported with higher confidence than details originating from the initial event, producing overconfidence (Flowe et al., 2019; Spearing & Wade, in press).

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

Consistent with previous research, we found that people's beliefs about their general memory ability did not significantly predict their confidence when they recalled details shortly after witnessing a mock crime event (Saraiva et al., 2020). We predicted that self-rated memory ability would have a larger impact on confidence when the delay was long compared to when the delay was relatively short. We found that people gave lower confidence judgements after longer delays, yet their beliefs about their general memory ability did not affect this pattern. One explanation for these findings is that the metamemory questionnaires focused on participants' beliefs about their everyday memory ability, and these beliefs may not extend to eyewitness scenarios. Developing new measures that specifically focus on eyewitness recall may help to provide further insight on the impact of

self-rated memory ability on witness's confidence judgements. Another explanation is that people may recognise that they have little experience of reporting details in an eyewitness context and therefore rely more on theory-based cues arising from the witnessing or testing situation (e.g., lighting conditions), or experience-based cues that arise from the process of remembering (e.g., retrieval fluency, Koriat, 1997).

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

Finally, future research could examine how long delays affect the specificity of witnesses' responses. We know that general responses (e.g., "dark") tend to be more accurate than specific responses (e.g., "dark blue," Sauer & Hope, 2016), yet witnesses rarely give general responses in their open-ended reports (Brewer, Vagadia, Hope & Gabbert, 2018). Thus, one possibility is that instructing witnesses to vary the specificity of their responses may help to maintain accuracy over long delays. However, is it important to note that general responses tend to produce poorer calibration and greater underconfidence than specific

responses (Spearing & Wade, in press) so encouraging witnesses to report more general responses may enhance memory but at the cost of impairing the confidence-accuracy relationship.

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

Author Contributions

Both authors were involved in the design of the experiments and the drafting of the manuscript. The first author performed data collection and conducted statistical analyses. Both authors have read and approved the final version.

Funding

The first author was supported by a University of Warwick Departmental PhD Fellowship.

Acknowledgements

We are grateful to Deryn Strange for supporting data collection. We would also like to thank Lorraine Hope and two anonymous reviewers for many helpful suggestions during the review process.

References

Ames, D. R., Rose, P., & Anderson, C. P. (2006). The NPI-16 as a short measure of narcissism. *Journal of Research in Personality*, *40*, 440-450.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*, 289-300.

Berkowitz, S. R., Frenda, S. J. (2018). Rethinking the Confident Eyewitness: A Reply to Wixted, Mickes, and Fisher. *Perspectives on Psychological Science*, *13*, 336-338.

Bradfield, A. L., & Wells, G. L. (2000). The perceived validity of eyewitness identification testimony: A test of the five Biggers criteria. *Law and Human Behavior*, *24*, 581-594.

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

Brewer, N., & Weber, N. (2008). Eyewitness confidence and latency: Indices of memory processes not just markers of accuracy. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, *22*, 827-840.

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.

Charman, S. D., Carlucci, M., Vallano, J., & Gregory, A. H. (2010). The selective cue integration framework: A theory of postidentification witness confidence assessment. *Journal of Experimental Psychology: Applied*, *16,* 204.

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

Cutler, B. L., Penrod, S. D., & Stuve, T. E. (1988). Juror decision making in eyewitness identification cases. *Law and Human Behavior*, *12*, 41-55.

Desmarais, S. L., & Read, J. D. (2011). After 30 years, what do we know about what jurors know? A meta-analytic review of lay knowledge regarding eyewitness factors. *Law and Human Behavior*, *35,* 200-210.

Evans, J. R., & Fisher, R. P. (2011). Eyewitness memory: Balancing the accuracy, precision and quantity of information through metacognitive monitoring and control. *Applied Cognitive Psychology, 25,* 501-508.

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

Fox, S. G., & Walters, H. A. (1986). The impact of general versus specific expert testimony and eyewitness confidence upon mock juror judgment. *Law and Human Behavior*, *10,* 215-228.

Gabbert, F., Hope, L., & Fisher, R. P. (2009). Protecting eyewitness evidence: Examining the efficacy of a self-administered interview tool. *Law and Human Behavior*, *33*, 298-307.

Garrett, B. L., Liu, A., Kafadar, K., Yaffe, J., & Dodson, C. S. (2020). Factoring the Role of Eyewitness Evidence in the Courtroom. *Journal of Empirical Legal Studies*, *17,* 556-579.

Goldsmith, M., Koriat, A., & Weinberg-Eliezer, A. (2002). Strategic regulation of grain size memory reporting. *Journal of Experimental Psychology: General, 131,* 73–95.

Horry, R., Colton, L. M., & Williamson, P. (2014). Confidence–accuracy resolution in the misinformation paradigm is influenced by the availability of source cues. *Acta Psychologica*, *151*, 164-173.

Iida, R., Itsukusima, Y., & Mah, E. Y. (2020). How do we judge our confidence? Differential effects of meta-memory feedback on eyewitness accuracy and confidence. *Applied Cognitive Psychology*, *34,* 397-408.

Jackson, S. A., & Kleitman, S. (2014). Individual differences in decision-making and confidence: capturing decision tendencies in a fictitious medical test. *Metacognition and Learning*, *9*, 25-49.

Juslin, P., Olsson, N., & Winman, A. (1996). Calibration and diagnosticity of confidence in eyewitness identification: Comments on what can be inferred from the low confidence–accuracy correlation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 1304-1316.

Kleitman, S., Hui, J. S. W., & Jiang, Y. (2019). Confidence to spare: individual differences in cognitive and metacognitive arrogance and competence. *Metacognition and Learning*, *14*, 479-508.

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.

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.

Lindholm, T., Jönsson, F. U., & Liuzza, M. T. (2018). Retrieval effort cues predict eyewitness accuracy. *Journal of Experimental Psychology*: *Applied, 24*, 534-542.

Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime. com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods*, *49*, 433-442.

Odinot, G., & Wolters, G. (2006). Repeated recall, retention interval and the accuracy–confidence relation in eyewitness memory. *Applied Cognitive Psychology*, *20*, 973-985.

Odinot, G., Wolters, G., & van Giezen, A. (2013). Accuracy, confidence and consistency in repeated recall of events. *Psychology, Crime & Law*, *19*, 629-642.

Ost, J., Easton, S., Hope, L., French, C. C., & Wright, D. B. (2017). Latent variables underlying the memory beliefs of Chartered Clinical Psychologists, Hypnotherapists and undergraduate students. *Memory*, *25*, 57-68.

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.

Pezdek, K., Abed, E., & Reisberg, D. (2020). Marijuana impairs the accuracy of eyewitness memory and the confidence–accuracy relationship Too. *Journal of Applied Research in Memory and Cognition*, *9*, 60-67.

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

Robinson, M. D., Johnson, J. T., & Herndon, F. (1997). Reaction time and assessments of cognitive effort as predictors of eyewitness memory accuracy and confidence. *Journal of Applied Psychology*, *82*, 416-425.

Saraiva, R. B., Hope, L., Horselenberg, R., Ost, J., Sauer, J. D., & van Koppen, P. J. (2020). Using metamemory measures and memory tests to estimate eyewitness free recall performance. *Memory*, *28,* 94-106.

Sauerland, M., Broers, N. J., & Van Oorsouw, K. I. M. (2019). Two field studies on the effects of alcohol on eyewitness identification, confidence, and decision times. *Applied Cognitive Psychology*, *33*, 370-385.

Sauerland, M., & Sporer, S. L. (2009). Fast and confident: postdicting eyewitness identification accuracy in a field study. *Journal of Experimental Psychology: Applied*, *15*, 46-62.

Semmler, C., Brewer, N., & Wells, G. L. (2004). Effects of postidentification feedback on eyewitness identification and nonidentification confidence. *Journal of Applied Psychology*, *89*, 334-346.

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

Shaw, J. S., & McClure, K. A. (1996). Repeated postevent questioning can lead to elevated levels of eyewitness confidence. *Law and Human Behavior*, *20*, 629-653.

Spearing, E. R., & Wade, K. A. (in press). Providing Eyewitness Confidence Judgements During Versus After Eyewitness Interviews Does Not Affect the Confidence-Accuracy Relationship. *Journal of Applied Research in Memory and Cognition.*

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.

Troyer, A. K., & Rich, J. B. (2002). Psychometric properties of a new metamemory questionnaire for older adults. *Journals of Gerontology: Psychological Sciences*, *57,* 19-27.

Tuckey, M. R., & Brewer, N. (2003). The influence of schemas, stimulus ambiguity, and interview schedule on eyewitness memory over time. *Journal of Experimental Psychology: Applied*, *9*, 101.

Wade K. A., Nash R.A., & Lindsay, D.S. (2018). Reasons to Doubt the Reliability of Eyewitness Memory: Commentary on Wixted, Mickes, and Fisher (2018). *Perspectives on Psychological Science*. *13*, 339-342.

Wixted, J. T., Mickes, L., & Fisher, R. P. (2018). Rethinking the reliability of eyewitness memory. *Perspectives on Psychological Science*, *13*, 324-335.

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.

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