***** ACCEPTED FOR PUBLICATION IN 'JOURNAL OF APPLIED RESEARCH IN MEMORY AND COGNITION' ON APRIL 22, 2021 *****

Providing Eyewitness Confidence Judgements During Versus After Eyewitness Interviews Does Not Affect the Confidence-Accuracy Relationship

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

Recent studies suggest that highly confident eyewitnesses are likely to provide highly accurate identification evidence, at least in some conditions. Yet few studies have investigated the confidence-accuracy relationship in witness interviews or exactly when confidence judgements should be taken. Across three experiments, 831 adults answered questions about a mock crime and rated their confidence in each response. Participants gave their confidence immediately after each response or at the end of the memory test. The timing of the confidence judgement did not affect the confidence-accuracy relationship, and the confidence-accuracy relationship remained strong even when participants encoded the event under poor visibility conditions. When participants were unknowingly exposed to misinformation, however, the confidence-accuracy relationship was substantially weakened—participants became highly over-confident in the accuracy of their memories. These findings help to refine the parameters in which witness confidence serves as a useful indicator of memory accuracy.

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

General Audience Summary

A growing body of research suggests that highly confident eyewitnesses are likely to be highly accurate, at least under some conditions. Yet most of what memory scientists know about the relationship between witness confidence and accuracy comes from eyewitness identification studies in which people watch a mock crime before attempting to identify a perpetrator from a lineup. Few studies have investigated the relationship between the accuracy of eyewitnesses' memory reports for crime-related details (e.g., the perpetrator's features, actions etc) and their confidence in these reports. Furthermore, there is no consensus on when confidence judgements should be elicited from witnesses during witness interviews. In three experiments, we investigated whether the relationship between confidence and accuracy is affected by when eyewitnesses give their confidence judgements. Participants watched a mock crime video and then completed a memory test for the event. Some participants gave their confidence judgements immediately after each response, whereas other participants only gave their confidence judgements after they had reported everything that they could remember. The proportion of details correctly remembered increased with increasing levels of confidence, regardless of when these confidence judgements were taken. In addition, people appropriately reduced their confidence if their memory was reduced by poor visibility, but not when they were unknowingly exposed to misinformation. These findings suggest that eliciting confidence judgements from witnesses at the end of their statements may not reduce their informative value any more than taking confidence judgements immediately after each response. Our data show, however, that people may become over-confident when they fail to notice factors that reduce their memory accuracy.

Introduction

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

Yet, some memory experts continue to warn about the risks of overstating the value of witness confidence in criminal investigations. They have posited that we do not have enough data to fully understand the factors that influence the confidence-accuracy relationship in real cases (Berkowitz, Garrett, Fenn, & Loftus, 2020; Berkowitz & Frenda, 2018; Sauer, Palmer & Brewer, 2019; Wade, Nash & Lindsay, 2018). Furthermore, most of what we know about eyewitness confidence and accuracy is based on eyewitness identification research—studies in which people make a single decision in an attempt to identify a perpetrator from a lineup (e.g., Colloff, Wade, Wixted, & Maylor, 2017; Palmer, Brewer, Weber & Nagesh, 2013). Relatively few studies have explored the confidence-accuracy relationship in witness

interviews, where mock-witnesses are asked to recall crucial details, such as a perpetrator's features, clothing and actions. In the current study, we investigated the confidence-accuracy relationship in witness interviews and gathered data on a factor that might impair this association: The timing of confidence ratings.

The few studies that have explored the confidence-accuracy relationship in witness interviews suggest that confidence may be a good proxy for accuracy in some contexts (Odinot, Wolters & van Giezen, 2013; Weber & Brewer, 2008). For instance, a recent study found that confidence was strongly related to accuracy, regardless of whether details were elicited through open-ended interview instructions (i.e., "Write down everything you can remember") or forced report cued-recall questions (e.g., "What was the hair colour of the customer?") (Brewer, Vagadia, Hope, & Gabbert, 2018). Other studies have shown that the confidence-accuracy relationship remains strong even when memory is weak because people tend to adjust their confidence when they recognise that their memory has been compromised (Koriat, 1997). For example, participants have appropriately lowered their confidence to compensate for their lower accuracy after their attention was divided between encoding and a secondary task, and when the delay between encoding and the memory test was increased (Odinot & Wolters, 2006; Sauer & Hope, 2016). Taken together, this research suggests that confidence can be informative about the accuracy of eyewitness reports at least in some circumstances.

There are circumstances, however, in which the confidence-accuracy relationship breaks down. A growing number of studies reveal that when people are unknowingly exposed to misinformation, they often report this information with high confidence (Flowe et al., 2019; Hope, Gabbert, Healey, & Lenton, 2008; Wright, Self & Justice, 2000). Confidence can also be inflated by factors that do not affect memory accuracy, such as when people receive positive feedback about their memory ability (e.g., "you're spot on"; Iida, Itsukusima & Mah,

2020) or answer easy questions before difficult questions in a test (Michael & Garry, 2019). These studies suggest that the context in which witnesses' confidence and memory reports are gathered are vital to preserving a strong confidence-accuracy relationship.

One key testing condition, however, that has been largely ignored in the eyewitness literature, is the timing of witnesses' confidence judgements. A review of published research shows that some studies have collected confidence immediately after participants reported each detail (*immediate-confidence* judgements; e.g., Dodson & Krueger, 2006; Paulo, Albuquerque & Bull, 2016), whereas others have collected confidence only after participants reported everything they could remember (*delayed-confidence* judgements; e.g., Evans & Fisher, 2011; Roberts & Higham, 2002). Given that the methodologies differed substantially across these studies—including the nature of the stimuli, the retention intervals, the interview procedures and testing formats—it is difficult to ascertain how the timing of confidence judgements might have influenced the confidence-accuracy relationship.

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

same viewing conditions—so the resulting confidence-accuracy correlation almost certainly underestimates the correlation that would be observed in real-world witnessing situations (Lindsay, Nilsen & Read, 2000; Lindsay, Read & Sharma,1998).

There are at least two reasons to predict that the timing of confidence judgements might influence the confidence-accuracy relationship. Previous work suggests witnesses' confidence could become inflated when their confidence judgements are delayed. Many studies show that repeated exposure to information can increase the ease with which people process that information, which in turn increases the subjective impression that the information is accurate (Alter & Oppenheimer, 2009; Kelley & Lindsay, 1993; Unkelbach & Stahl, 2009). People report, for instance, being more confident they have visited a specific location, like a university campus, after viewing a photo of that location twice rather than only once (Brown & Marsh, 2008). People are also more confident that an eyewitness' erroneous claims are true after reading those claims three times rather than once (Foster, Huthwaite, Yesberg, & Garry, 2012). One important distinction between immediate and delayed-confidence judgements is that when confidence judgements are made immediately, witnesses only consider their responses at the point of retrieval. But when confidence judgements are delayed, witnesses consider their responses at the point of retrieval and then again, after a delay, when they are asked to provide confidence judgments for each detail they have recalled. In short, when confidence judgements are delayed, witnesses are repeatedly exposed to their responses, which could enhance processing fluency and lead to over-confidence. If so, we might expect immediate-confidence participants to show stronger confidence-accuracy calibration than delayed-confidence participants. Unpublished lineup research supports this prediction, showing that confidence judgements taken immediately after an identification produced stronger calibration than confidence judgements taken after a short 5-minute delay (Brewer, Weber & Semmler, 2005).

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

In three experiments, we examined whether the timing of confidence judgments affects the confidence-accuracy relationship in eyewitness interviews. In each experiment, participants watched a mock crime video, and after a delay completed a memory test and rated their confidence in their responses. Some participants rated their confidence immediately after providing each response, and others rated their confidence at the end of the memory test. Each experiment was preregistered; the numeric data for all experiments, and the corresponding R code are available on Open Science Framework: https://osf.io/mp3r8/?view_only=5c581f7fd5974409b1d3877bf48158a2 for Experiment 1, https://osf.io/gqkyp/?view_only=4e1be3b55ff3412ea09e203589279099 for Experiment 2, and https://osf.io/dbmnc/?view_only=59277e823a27491c8778185ef10d4cd2 for Experiment 3.

Experiment 1

Method

Participants & Design

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

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

Materials

Videos. We created two mock crime videos for the experiment, a car theft and a mugging scenario. In the car theft scenario (2 min 37 s), a female thief walks around a supermarket car park and peeks into a parked car. She notices the female victim leaving her car and goes to steal it. The victim sees the thief driving away and chases after the car. In the mugging scenario (3 min 16 s), a female victim exchanges phone numbers with a male friend. She puts the phone into her bag and a male thief instructs her to give him the bag. When she refuses, he wrestles the bag from her and runs away. Using Adobe After Effects®, we digitally altered the two original videos so they resembled a night scene, producing a total of four videos (see Figure 1). The stimuli were pilot tested to ensure that the night videos produced a lower level of memory accuracy than the day videos whilst avoiding floor effects.

Figure 1. Screenshots of the car theft (left) and mugging (right) videos with day (top) and night visibility (bottom).

Memory tests. We created a memory test for each crime scenario that contained 18 4-alternative forced choice (4AFC) questions (see the supplementary materials for the full memory test). Questions pertained to people, actions, objects, and locations in the video (e.g.,

"What was the colour of the stolen bag?"). Each question was presented with a correct answer and three incorrect alternatives (e.g., correct answer: blue, incorrect alternatives: grey, black, brown).

Procedure

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

In Phase 2, participants completed the surprise 4AFC memory test. They did not receive any feedback on their memory performance. Participants rated their confidence on a scale from 25% (not at all confident) to 100% (very confident) that increased in increments of 5%. They were told that 25% was the level of accuracy expected from guessing. We used a 25-100% scale to reflect the minimum to maximum expected frequency of a correct response on a 4AFC memory test (Bornstein & Zickafoose, 1999; Tekin & Roediger, 2017). This enabled us to analyse the correspondence between the expected and observed frequencies of a correct response with calibration statistics. Immediate-confidence participants rated their confidence immediately after answering each question whereas delayed-confidence participants rated their confidence after completing the entire memory test. Delayed-confidence participants were reminded of their responses as they rated their confidence for each question (e.g., "Question: What was the colour of the stolen bag? Your response: Blue. How confident are you that your response is correct?"). Participants were then asked if they experienced any technical difficulties while watching the video, which device they were

using, and if they had complied with the criteria outlined in Phase 1. Finally, participants answered some demographic questions and were debriefed.

Results

Preliminary Analyses

Before turning to our main analyses, we conducted two preliminary analyses. First, we checked whether calibration was similar across the two events by plotting calibration separately for the mugging and car theft scenarios. Calibration plots were created by plotting accuracy (i.e., the proportion correct) against four levels of confidence (25 – 40, 45 – 60, 65 – 80, and 85 – 100). Table S10 in the supplementary materials shows the count data for each confidence bin (and includes count data for the confidence bins all three experiments). The calibration curve indicated that calibration did not significantly differ between the two events (see Figure S1 in the supplementary materials). Second, we checked that memory accuracy was lower in the night condition (M = 60.92; 95% CI = 59.00, 62.83) than in the day condition (M = 67.86; 95% CI = 66.14, 69.57), and the non-overlapping CIs indicate that it was.

Main Analyses

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

To answer our key question, we conducted a mixed binary logistic regression on response accuracy (i.e., correct vs incorrect) with confidence, confidence timing and visibility

as predictors and a random intercept for each participant. This method is preferred to ANOVAs for analysing categorical data, as the latter can yield spurious results (see Jaeger, 2008). Additionally, the inclusion of random intercepts controls for unmeasured variables that are related to memory performance (e.g., fatigue). The model was fitted using the *lme4* package v1.1-23 in R 4.0.2 (Bates, Maechler, Bolker, & Walker, 2015) and p-values were obtained in the *afex* package v. 0.26-1 (Singmann, Bolker, Westfall, & Aust, 2020). In line with existing research (e.g., Saraiva et al., 2020; Weber & Brewer, 2008; Wixted & Wells, 2017), the analyses revealed that confidence was a significant predictor of accuracy, $\chi^2(1) = 1003.30$, $p < .001$. The odds ratio was 2.61, indicating that each unit increase in confidence more than doubled the likelihood of a correct answer.

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

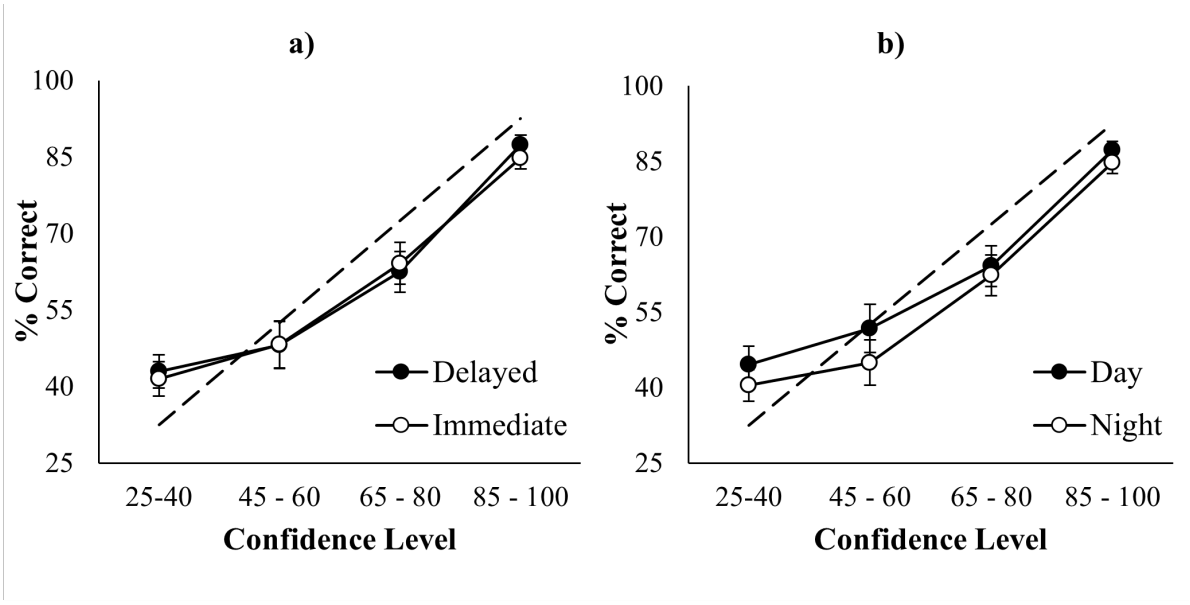

Figure 2. Calibration plots for a) confidence timing and b) visibility, with the dashed line representing perfect calibration. Error bars denote the 95% CI around the mean.

Finally, as Figure 2b shows, the visibility (day vs night scene) manipulation did not affect the strength of the confidence-accuracy relationship. The full visibility analyses are available in the supplementary materials.

In Experiment 1, the confidence-accuracy relationship was strong when participants were asked relatively innocuous questions, regardless of the timing of the confidence judgements. In Experiment 2, we aimed to replicate our findings in a cued recall test format and to investigate how misleading questions affect the confidence-accuracy relationship. Decades of research shows witnesses can report misinformation they have gleaned from several sources, and sometimes report it with high confidence. If participants fail to notice the discrepancy between the target event and the misleading post-event information, then they may use retrieval fluency at test to judge their confidence (Horry, Colton & Williamson, 2014; Lindsay & Johnson, 1989). Relying on retrieval fluency can serve to inflate a person's confidence because misinformation is encountered more recently than the original event, and as such, should be recalled with more ease. If so, participants should show more over-

confidence and poorer calibration when they have been misled about an item than when they have not been misled.

Experiment 2

Method

Participants & Design

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

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

Procedure

Participants completed the three-phase study online and were randomly assigned to one of the 4 between-participant groups. They were told that the study was about "perception of events" and asked to comply with several criteria during the experiment (e.g., "please complete the experiment in a single sitting, and do not stop the experiment to complete other

tasks"). In Phase 1 (target event), participants watched either the car theft or mugging video (both with day visibility). We selected 12 critical items from each event. For example, critical items in the car theft video included the *name of the parking zone*, whereas critical items in the mugging video included the *colour of the thief's trousers*. The video was followed by two attention check questions (e.g., "Was the victim male or female?") and a 3-minute filler task.

In Phase 2 (misinformation), participants were randomly assigned to answer one of three questionnaires containing 8 x 2AFC questions each referring to a different critical item. They were told that each question had two possible responses, and that they should select a response for each question. Four questions were misleading and contained two incorrect answers (e.g., "*Was the stolen car parked in Parking Zone 4 or Parking Zone D*?", correct response = Parking Zone 3). The remaining four questions were not misleading and contained the correct answer and one incorrect answer (e.g., *"Was the stolen car parked in Parking Zone 3 or Parking Zone D?"*). The critical items were counterbalanced so that – across the three questionnaires – each item appeared once as a misled item and once as a consistent item. That is, a third of participants answered the misleading question about the parking zone, another third answered the consistent question about the parking zone, and the remaining participants did not answer a question about the parking zone. After answering the questionnaire, participants completed a 2.5-minute filler task.

In Phase 3 (memory test), all participants completed the same 12-item cued recall test with one question for each of the 12 critical items (e.g., "*According to the sign, which parking zone was the stolen car parked in?",* see the supplementary materials for the full memory test*).* Eight questions referred to critical items that participants were previously asked about in the misleading questionnaire (4 misled items, 4 consistent items), and 4 questions referred to control items that were not asked about in the misleading questionnaire. Participants were informed that some questions related to details that they were asked about

previously and told that they should respond based on their memory of the video, and not on their memory of their previous responses. Participants rated their confidence on an 11-point scale, ranging from 0% (not at all confident) to 100% (very confident). Immediate-confidence participants rated their confidence immediately after answering each question. Delayed-confidence participants rated their confidence after completing the entire memory test. They were reminded of their responses as they rated their confidence for each question (e.g., "Question: Where did the thief put their phone? Your response: In their pocket. How confident are you in your response?"). Participants were then asked (1) if they experienced any technical difficulties while watching the video, (2) which device they were using, (3) if they had complied with the criteria outlined in Phase 1 and (4) if they had any suspicions about the true purpose of the experiment. Finally, participants answered some demographic questions and were fully debriefed.

Results

Preliminary Analyses

As in Experiment 1, we checked that calibration was similar for the car theft and mugging scenarios. We created calibration curves by plotting accuracy against 5 levels of confidence (0 – 20, 30 – 40, 50 – 60, 70 – 80 and 90 – 100). "I don't know" responses (n = 234, 7% of responses) were excluded and not analysed. Therefore, accuracy was calculated as the number of correct responses divided by the number of correct and incorrect responses (total n = 3,126). The calibration curve revealed that accuracy was lower for the mugging video than for the car theft video, but the overall pattern of calibration was similar for the two events (see Figure S2 the supplementary materials). We also checked that accuracy was lower for misled items (M = 36.99; CI = 33.86, 40.13) than for consistent (M = 75.82, CI = 72.93,

78.68) and control items (M = 59.32, CI = 55.96, 62.50), which it was. Accuracy was also significantly higher for consistent items than control items.

Main Analyses

Recall that our main aims were to replicate the findings for confidence timing and to investigate the influence of misinformation on the confidence-accuracy relationship. To address these aims, we collapsed the data across events then conducted a binary logistic regression on response accuracy (correct vs incorrect) as in Experiment 1, with confidence, confidence timing, and item type as predictors. The model confirmed that confidence was a significant predictor of accuracy, $\chi^2(1) = 392.97$, $p < .001$. The odds ratio was 2.57, very similar to Experiment 1 (2.61), indicating that each point increase in confidence doubled the odds of a correct response. Consistent with Experiment 1, the regression model revealed that there was no significant effect of confidence timing, $\chi^2(1) = 0.01$, $p = .91$, nor did confidence timing significantly interact with confidence, $\chi^2(1) = 0.52$, $p = .47$. Furthermore, Figure 3a shows that immediate and delayed-confidence judgements produced similar calibration at every level of confidence. Thus, the time at which participants provided confidence judgements did not significantly influence the confidence-accuracy relationship for cued-recall.

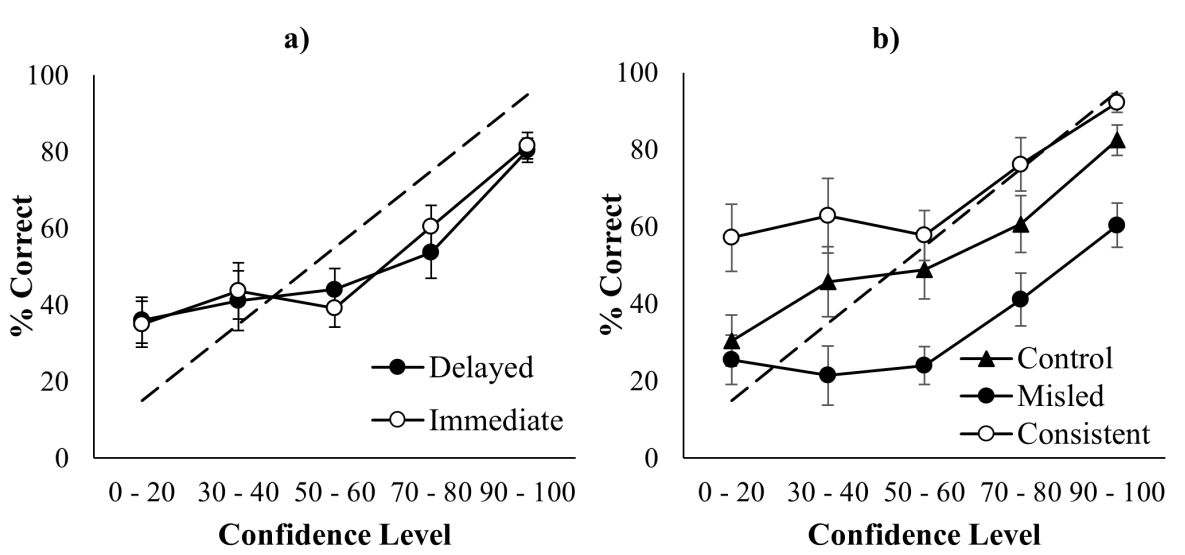

Figure 3. Calibration plots for a) confidence timing and b) item type. The dashed line represents perfect calibration, and the error bars denote the 95% CI around the mean.

Is the confidence-accuracy relationship weaker for misled items than for consistent and control items? The regression model revealed that item type was a significant predictor of accuracy, $\chi^2(2) = 304.03$, $p < .001$. Participants were 64% less likely to be correct for misled items (OR = 0.36) and 106% more likely to be correct for consistent items than control items (OR = 2.06), and this effect was stable across confidence levels $\chi^2(2) = 2.83$, $p = .24$.

To what extent did different item types deviate from perfect calibration? Figure 3b shows that misled items produced significant over-confidence at almost every level of confidence. Consistent items, in contrast, produced significant under-confidence at the two lowest levels of confidence, but the best calibration at higher levels of confidence. Consistent questioning may have strengthened participants' memory for the target items, which increased accuracy with a smaller effect on confidence. Finally, control items produced the most linear increase in accuracy across increasing levels of confidence, resulting in under-confidence at the lowest levels of confidence and over-confidence at the highest levels of confidence.

To summarise, in Experiments 1 and 2, we found that the timing of confidence judgements did not significantly influence the confidence-accuracy relationship for recognition or cued-recall tasks. Real eyewitnesses, however, are often subjected to lengthy interviews in which they are encouraged to freely report everything that they can remember. In Experiment 3, we aimed to replicate our findings in a more forensically relevant recall task. We also examined whether the timing of confidence judgements influences the completeness of eyewitnesses' memory reports. Asking participants for confidence judgements immediately after each response could disrupt their narration and the flow of recall (see Fisher & Geiselman, 2010). As a result, we might expect immediate-confidence participants to provide shorter reports (i.e., recall fewer details) than delayed-confidence participants.

To investigate this, participants watched one mock crime video, completed free recall and cued recall tests for details in the video, and then completed this procedure again for an alternative mock crime video. Research shows that eyewitness free recall reports are typically associated with high accuracy at every level of confidence, and responses are mostly given with high confidence (Saraiva et al., 2020). Therefore, we included cued-recall questions to produce enough variation in accuracy and confidence to plot stable calibration curves.

Experiment 3

Method

Participants & Design

We used a 2 (Event: car theft, mugging) x 2 (Confidence timing: immediate, delayed) mixed design with event as the within-participants factor. As people rarely report responses held with low confidence under free report conditions, this study required more observations to achieve stable calibration estimates. Therefore, we aimed to collect 10 observations from

at least 200 participants, producing >1,000 observations in each between-participants condition. In total, we collected data from 238 participants. Of these participants, 146 were first year Psychology students at the University of Warwick who participated in partial fulfilment of course requirements. The remaining 92 participants were recruited from the wider university community and received £6 upon completing the experiment. We excluded those who did not complete the recall tasks correctly (e.g., if they did not provide confidence ratings, n = 20) or reported technical difficulties (n = 7). The final sample consisted of 211 participants (43 male, 164 female, 4 undisclosed, M = 19.77 years, SD = 3.05, range = 18-40), including 111 in the immediate-confidence condition and 100 in the delayed-confidence condition.

Procedure

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

In Phase 2, participants completed the free-recall task. Instructions were presented on the computer screen and participants completed the test in a paper response booklet. Participants were instructed to write down everything that they could remember about the video and to write each detail on a new line. They were told that they could vary the level of detail in their responses and given examples of general-level responses (e.g., "there were 3 to 6 people in the shop") and specific-level responses (e.g., "there were 4 people in the shop"). The examples were unrelated to the content in the video and showed how specificity could be

varied for different types of details. Immediate-confidence participants rated their confidence from 0 (not at all confident) to 100 (very confident) immediately after writing each response. Delayed-confidence participants were asked to write down everything they could remember and to leave the box next to each response blank.

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

This two-phase procedure was then repeated for the alternate video. The order of the video was counterbalanced. Finally, participants were asked if they experienced any technical difficulties while watching either of the videos, answered some demographic questions and were fully debriefed.

Data Coding

Participants’ written responses were coded as correct, incorrect, or not applicable (“don’t know”). Following Vredeveldt and Sauer (2015), responses were coded as incorrect if they contained any incorrect details even if they were partly accurate. For example, if the thief put money into a *blue* backpack then “the thief put money in a *green* backpack” would be coded as incorrect. Responses were also coded for specificity. Specific answers included a precise description of an item, person or location in the video (e.g., “the thief wore blue skinny jeans”), whereas general answers included only a broad, imprecise description (e.g.,

"the thief wore jeans"). When there was no clear distinction between general and specific, responses were coded as not applicable. All responses were coded by two independent coders who were blind to participants' conditions, and percentage agreement exceeded 89% for accuracy ($M = 0.89$, $K = 71$, $p < .001$) and specificity ($M = 0.90$, $K = 0.84$, $p < .001$). The same raters then discussed discrepancies to reach total agreement.

Results

Preliminary Analyses

As in Experiments 1 and 2, we checked that calibration was similar for the car theft and mugging scenarios. Responses coded as NA for accuracy or confidence (i.e., when confidence could not be read) were removed from the analysis ($N = 1{,}160$). In total, 11,204 responses were included in our analysis. Accuracy was higher for the car theft video except at the lowest level of confidence, but the pattern of calibration was similar across events. Next, we checked that calibration was similar for the first and second video, which it was. There were no significant differences at any level of confidence. The calibration curves are provided in Figure S3 in the supplementary materials.

Main Analyses

Turning to our main research question: Is the confidence-accuracy relationship affected by the timing of confidence judgements in eyewitness recall? To answer this question, we collapsed the data over event type and video order, then conducted a binary logistic regression on response accuracy (correct vs incorrect) with confidence and confidence timing as predictors. To produce enough variation in accuracy and confidence to plot stable calibration curves, we aggregated each participant's responses in the cued recall and free recall tests. Mean accuracy scores and confidence ratings are presented separately for the cued recall and free recall tests in Table S9 in the supplementary materials. As in

Experiments 1 and 2, confidence was a significant predictor of accuracy, χ^2 (1) = 1332.30 (OR = 2.37), $p < .001$, such that the odds of a correct response doubled with each unit increase in confidence.

The regression model revealed, once again, that confidence timing did not significantly predict accuracy, χ^2 (1) = 3.49, $p = .06$. Due to the limited number of observations at low levels of confidence (0 – 20), we did not include a confidence x confidence timing interaction in the regression model. However, Figure 4 shows that calibration was similar between the two groups.

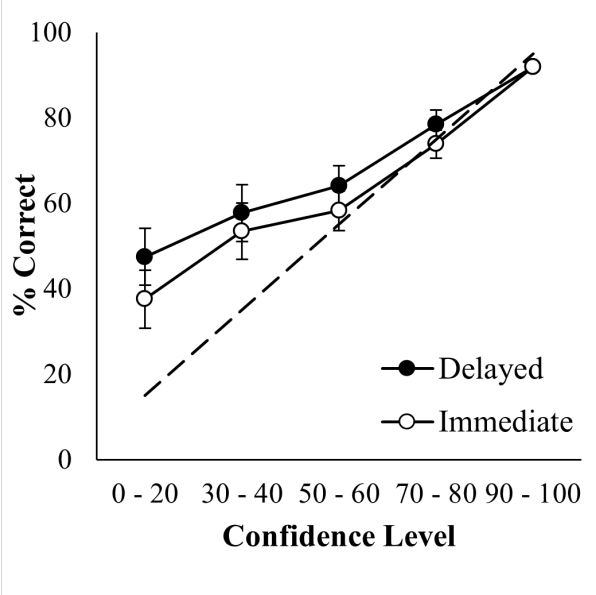

Figure 4. Calibration plot for immediate and delayed confidence judgements. Error bars denote the 95% CI around the mean.

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

participants' reports, F (1, 209) = 1.75 p = .19. For a summary of the free recall data, see the supplementary materials.

Exploratory Analysis

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

Several studies show that specific details are associated with greater over-confidence than general details, but less is known about how specificity affects' eyewitness calibration (Brewer et al., 2018; Vredeveldt & Sauer, 2015). Figure 5 shows that calibration was better for specific details than general details. In line with previous research, specific details produced strong calibration and only a small amount of overconfidence at the highest level of confidence. General details, however, produced poorer calibration and significant underconfidence at every level of confidence except the highest level.

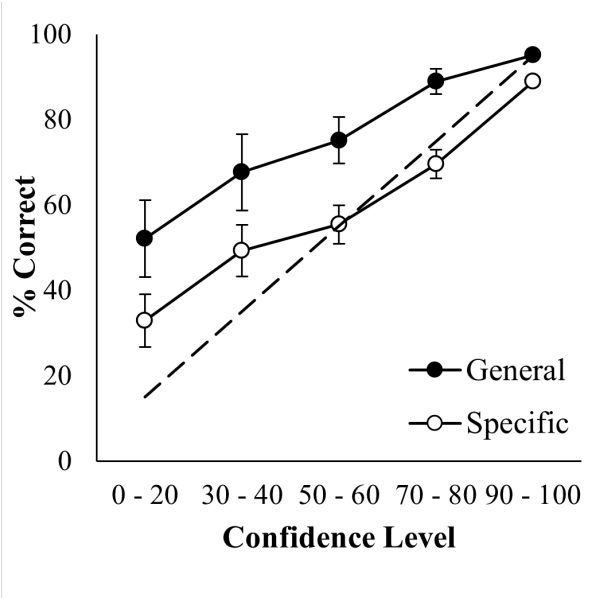

Figure 5. Calibration plot for general and specific responses. Error bars denote the 95% CI around the mean.

Discussion

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

Given there were good reasons to predict that delaying witnesses' confidence judgements should impair the confidence-accuracy relationship, why did we not observe any effect of confidence timing here? One possible explanation, guided by Koriat's cue-

utilization theory (Koriat, 1997), is that people use a variety of memorial cues to make their confidence responses, including their own metamemorial beliefs. That is, people have common-sense—but not necessarily accurate—beliefs about what factors do and do not affect memory accuracy, and people use these beliefs to decide whether to be relatively cautious or not when making confidence judgements. The relationship between confidence and accuracy may be disrupted when witnesses' fail to realise that their memory has been affected by a given factor, and as such, fail to appropriately adjust their confidence ratings. This mechanism could explain why delaying participants' confidence ratings did not affect confidence-accuracy calibration. We predicted that delayed-confidence participants might experience increased processing fluency, because they reviewed their memory responses—a form of re-exposure—while providing confidence ratings, and this may serve to inflate their confidence judgements. It is possible, however, that delayed-confidence participants correctly recognised the source of that increased fluency (indeed, we reminded them that they were viewing their own answers once again) and therefore deliberately made the appropriate adjustments in confidence.

The accumulating data on confidence-accuracy calibration in witness interviews appears to fit with this metamemorial account. Factors that have been shown to disrupt calibration such as exposure misinformation (Flowe et al., 2019) and marijuana intoxication (Pezdek, Abed & Reisberg, 2020) may be factors for which people tend to hold misguided metamemorial beliefs. For instance, a witness might not realize they have been exposed to misinformation or they may underestimate the impact of misinformation on memory and therefore see no reason to adjust their confidence. Conversely other factors have been shown to have little or no impact on calibration, including weapon focus (Carlson, Dias, Weatherford & Carlson, 2017), stress (Pezdek, Abed & Cormia, 2020), and attention (Sauer & Hope, 2016). These may be factors for which people tend to hold more accurate

metamemorial beliefs. We should note that we did not set out to test underlying mechanisms here, but a greater theoretical understanding of calibration in witness interviews will be essential to advancing procedures that maximise calibration, and to predicting a priori when the confidence-accuracy relationship is likely to be impaired.

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

Consistent with previous research, our data showed that factors that impair memory accuracy do not necessarily reduce the confidence-accuracy relationship (e.g., Palmer et al., 2013; Wixted & Wells, 2017). In Experiment 1, night visibility reduced the accuracy of

eyewitness reports, but did not compromise the confidence-accuracy relationship. Participants seemingly reduced their confidence to compensate for their lower memory accuracy, so calibration remained strong. This finding is not so surprising given that people are likely to be well aware of the effect of poor visibility on memory accuracy. Consistent with the metamorial account given above, when watching the event under low visibility (night) conditions, participants may have found it more difficult to make out precise details, such as the colour of objects, and subsequently interpreted this reduced processing fluency as a sign that their memory was not highly accurate. Additionally, participants may have appropriately reduced their confidence by applying a specific theory about how visibility affects memory performance (Busey, Tunnicliff, Loftus, & Loftus, 2000; Leippe, Eisenstadt & Rauch, 2009). While we cannot distinguish the basis for participants' confidence ratings in this study, our findings highlight that poor viewing conditions do not necessarily make for an unreliable witness as relatively high accuracy can still be observed at high levels of confidence (Wixted, Mickes & Fisher, 2018).

Whereas participants seemed to adapt their confidence judgements appropriately under different encoding conditions, they failed to do so when they encountered misleading questions. Decades of research shows that people often report information that is inconsistent with what they experienced, but studies have rarely investigated how misinformation affects calibration (Bonham & Gonzalez-Vallejo, 2009; Hope et al., 2008; Loftus, Miller & Burns, 1978; Morgan et al., 2013). When witnesses report misinformation with low confidence in the courtroom, this information is likely to be disregarded by triers of fact who use confidence to gauge the accuracy of a witness' testimony. As such, the misinformation is unlikely to result in a miscarriage of justice. Concerningly, we found that when participants were exposed to misinformation, they showed over-confidence in their memory accuracy at almost every level of confidence. These findings highlight concerns about the reliability of

eyewitness memory and using confidence ratings in forensic contexts where eyewitnesses may unknowingly encounter misinformation (Wade et al., 2018). Given that it may be impossible to determine when witnesses have been exposed to misinformation, our results suggest that interviewers should exercise caution when using confidence ratings for assessing the accuracy of eyewitness reports. They also raise further questions about how the confidence-accuracy relationship is affected by misinformation. For instance, how does misinformation gleaned from a co-witness versus an investigative officer (who is deemed to be authoritative) affect the confidence with which people report misinformation? And can calibration be preserved by warning witnesses that their memories may contain information from numerous sources?

Understanding the mechanisms behind witnesses' over-confidence may help researchers to better predict when the confidence-accuracy relationship will break down. One possibility is that misleading questions produced over-confidence because participants did not scrutinise the source of their memories and instead relied on retrieval fluency—that is, the speed with which information came to mind—to make their confidence judgement (Horry et al., 2014). Previous work suggests that retrieval fluency provides the rememberer with a good predictor of accuracy under most conditions, but it is deceptive when people are exposed to misinformation (Koriat, 1997). Misinformation is usually encountered after the to-be-remembered stimuli, so it tends to come to mind more quickly than the originally encoded information giving rise to the mistaken impression that the memory is accurate.

Finally, we found evidence to suggest that repeated questioning can improve memory accuracy in specific contexts (e.g., Odinot, Wolters & Lavender, 2009). Specifically, participants showed higher levels of memory accuracy on consistent items than on misled and control items. This finding fits with a growing body of research which suggests that repeated questioning does not produce confidence inflation, as originally thought (Odinot & Wolters,

2006; Odinot et al., 2013; Shaw & McClure, 1996). The novel contribution of the current studies, however, is that they show how repeated questioning influences eyewitness calibration. Specifically, we found that participants were more likely to show under-confidence on consistent items than on control items. Furthermore, participants maintained relatively high accuracy and strong calibration at high levels of confidence regardless of whether they were asked about an item once or twice.

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

Author Contributions

Both authors were involved in the design of the experiments and the drafting of the manuscript. The first author created the stimuli, 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 Alyssa Rashid and Georgina Carrigan for data coding, and to Samantha Rasor for help with data collection. We also thank Maryanne Garry and the Garry labsters for their thorough and insightful feedback on an earlier draft, and Dan Reisberg plus two anonymous reviewers for many excellent suggestions during the review process.

References

Alter, A. L., & Oppenheimer, D. M. (2009). Uniting the tribes of fluency to form a metacognitive nation. *Personality and Social Psychology Review*, *13*, 219-235.

Bates D, Mächler M, Bolker B, Walker S (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*, 1–48.

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

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.

Bonham, A. J., & González-Vallejo, C. (2009). Assessment of calibration for reconstructed eye-witness memories. *Acta Psychologica*, *131*, 34-52.

Bornstein, B. H., & Zickafoose, D. J. (1999). "I know I know it, I know I saw it:" The stability of the confidence–accuracy relationship across domains. *Journal of Experimental Psychology: Applied*, *5*, 76-88.

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

Brewer, N., Keast, A., & Sauer, J. D. (2010). Children's eyewitness identification performance: Effects of a Not Sure response option and accuracy motivation. *Legal and Criminological Psychology*, *15*, 261-277.

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., & Semmler, C. (2005). Eyewitness identification. In N. Brewer & K.D. Williams (Eds.), *Psychology and Law: An Empirical Perspective* (pp. 177–221). New York, NY: Guilford.

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.

Brown, A. S., & Marsh, E. J. (2008). Evoking false beliefs about autobiographical experience. *Psychonomic Bulletin & Review*, *15*, 186-190.

Busey, T. A., Tunnicliff, J., Loftus, G. R., & Loftus, E. F. (2000). Accounts of the confidence-accuracy relation in recognition memory. *Psychonomic Bulletin & Review*, *7*, 26-48.

Carlson, C. A., Dias, J. L., Weatherford, D. R., & Carlson, M. A. (2017). An investigation of the weapon focus effect and the confidence–accuracy relationship for eyewitness identification. *Journal of Applied Research in Memory and Cognition*, 6, 82-92.

Colloff, M. F., Wade, K. A., Wixted, J. T., & Maylor, E. A. (2017). A signal-detection analysis of eyewitness identification across the adult lifespan. *Psychology and Aging*, *32*, 243-258.

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

Deffenbacher, K. A. (1980). Eyewitness accuracy and confidence. *Law and Human Behavior*, *4*, 243-260.

Dodson, C. S., & Krueger, L. E. (2006). I misremember it well: Why older adults are unreliable eyewitnesses. *Psychonomic Bulletin & Review*, *13*, 770-775.

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.

Fisher, R. P., & Geiselman, R. E. (2010). The cognitive interview method of conducting police interviews: Eliciting extensive information and promoting Therapeutic Jurisprudence, *International Journal of Law and Psychiatry, 33,* 321-328.

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.

Foster, J. L., Huthwaite, T., Yesberg, J. A., Garry, M., & Loftus, E. F. (2012). Repetition, not number of sources, increases both susceptibility to misinformation and confidence in the accuracy of eyewitnesses. *Acta Psychologica*, *139*, 320-326.

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.

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

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

Hanway, P., Akehurst, L., Vernham, Z., & Hope, L. (2021). The effects of cognitive load during an investigative interviewing task on mock interviewers' recall of information. *Legal and Criminological Psychology*, *26*, 25-41.

Hope, L., Ost, J., Gabbert, F., Healey, S., & Lenton, E. (2008). "With a little help from my friends…": The role of co-witness relationship in susceptibility to misinformation. *Acta Psychologica*, *127*, 476-484.

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.

Huff, C. R. (1987). Wrongful conviction: societal tolerance of injustice. *Research in Social Problems and Public Policy, 4*, 99–115.

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.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*, 434-446.

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

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.

Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review*, *103*, 490-517.

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.

Lindsay, D. S., & Johnson, M. K. (1989). The eyewitness suggestibility effect and memory for source. *Memory & Cognition*, *17*, 349-358.

Lindsay, D. S., Nilsen, E., & Read, J. D. (2000). Witnessing-condition heterogeneity and witnesses' versus investigators' confidence in the accuracy of witnesses' identification decisions. *Law and Human Behavior*, *24*, 685-697.

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

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.

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.

Michael, R. B., & Garry, M. (2019). How do ordered questions bias eyewitnesses? *Memory*, *27*, 904-915.

Mickes, L. (2015). Receiver operating characteristic analysis and confidence–accuracy characteristic analysis in investigations of system variables and estimator variables that affect eyewitness memory. *Journal of Applied Research in Memory and Cognition*, *4*, 93-102.

Morgan III, C. A., Southwick, S., Steffian, G., Hazlett, G. A., & Loftus, E. F. (2013). Misinformation can influence memory for recently experienced, highly stressful events. *International Journal of Law and Psychiatry*, *36*, 11-17.

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., & Lavender, T. (2009). Repeated partial eyewitness questioning causes confidence inflation but not retrieval-induced forgetting. *Applied Cognitive Psychology*, *23*, 90-97.

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

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.

Paulo, R. M., Albuquerque, P. B., & Bull, R. (2016). Improving the enhanced cognitive interview with a new interview strategy: Category clustering recall. *Applied Cognitive Psychology*, *30*, 775-784.

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

Pezdek, K., Abed, E., & Cormia, A. (2020). Elevated stress impairs the accuracy of eyewitness memory but not the confidence–accuracy relationship. *Journal of Experimental Psychology: Applied*, 27, 158-169.

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.

Roberts, W. T., & Higham, P. A. (2002). Selecting accurate statements from the cognitive interview using confidence ratings. *Journal of Experimental Psychology: Applied*, *8*, 33-43.

Robinson, M. D., & Johnson, J. T. (1996). Recall memory, recognition memory, and the eyewitness confidence–accuracy correlation. *Journal of Applied Psychology*, *81*, 587-594.

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.

Sauer, J., & Hope, L. (2016). The effects of divided attention at study and reporting procedure on regulation and monitoring for episodic recall. *Acta Psychologica*, *169*, 143-156.

Sauer, J. D., Palmer, M. A., & Brewer, N. (2019). Pitfalls in using eyewitness confidence to diagnose the accuracy of an individual identification decision. *Psychology, Public Policy, and Law*, *25*, 147-165.

Scheck, B., Neufeld, P., & Dwyer, J. (2000). *Actual Innocence*. New York: Random House.

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.

Singmann, H., Bolker, B., Westfall, J., & Aust, F. (2018). afex': analysis of factorial experiments. R package.

Tekin, E., & Roediger, H. L. (2017). The range of confidence scales does not affect the relationship between confidence and accuracy in recognition memory. *Cognitive Research: Principles and Implications*, *2*, 1-13.

Unkelbach, C., & Stahl, C. (2009). A multinomial modeling approach to dissociate different components of the truth effect. *Consciousness and Cognition*, *18*, 22-38.

Vredeveldt, A., & Sauer, J. D. (2015). Effects of eye-closure on confidence-accuracy relations in eyewitness testimony. *Journal of Applied Research in Memory and Cognition*, *4*, 51-58.

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.

Weber, N., & Brewer, N. (2008). Eyewitness recall: Regulation of grain size and the role of confidence. *Journal of Experimental Psychology: Applied*, *14*, 50-60.

Wells, G. L., & Murray, M. (1984). Eyewitness confidence. In G. L. Wells, & E. F. Loftus (Eds.), *Eyewitness testimony: Psychological perspectives*, (pp. 155 – 170). New York: Cambridge University Press.

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.

Wright, D. B., Self, G., & Justice, C. (2000). Memory conformity: Exploring misinformation effects when presented by another person. *British Journal of Psychology*, *91*, 189-202.