Received: 23 November 2023 | Revised: 4 March 2024 | Accepted: 5 April 2024

DOI: 10.1002/acp.4197

RESEARCH ARTICLE

WILEY

Eyewitness confidence in the interviewing context: Understanding the impact of question type and order

Alessandra Caso [1] | Fiona Gabbert [2] | Coral J. Dando [1]

[1]Department of Psychology, School of Social Science, University of Westminster, London, UK

[2]Department of Psychology, Goldsmiths, University of London, London, UK

Correspondence
Alessandra Caso, Department of Psychology, School of Social Science, University of Westminster, London.
Email: a.caso2@westminster.ac.uk

Abstract

The relationship between confidence and accuracy and the reliability of eyewitness identifications has attracted a lot of attention. In contrast, relatively little is known about the relationship between eyewitness confidence and the accuracy of recall memory in interview contexts. Here, we manipulated questioning approaches to investigate the impact of Free-Recall and Cued-Recall questions, whereby the latter were witness-*compatible* (questions concerning details reported in the preceding Free-Recall) or witness-*incompatible* questions. We also manipulated the order these questions were asked. A sample of 124 mock witness participants watched a crime-video and subsequently recalled the event to understand the impact of question type and order on confidence-accuracy calibration. Our results show that a Free-Recall invitation and compatible (compared to incompatible) questions promoted more stable confidence. Compatible questions yielded fewer errors, more accurate details, and promoted more reliable confidence-accuracy calibration and discrimination, especially when they preceded the incompatible questions. Implications are discussed.

KEYWORDS
CA calibration, cued-recall questions, eyewitness confidence, interview question type, witness-compatible questions

1 | INTRODUCTION

Memory confidence is a term used to describe the extent to which individuals believe their memories to be accurate, for example when recognising a perpetrator in a line-up or when describing a crime in an interview. In forensic contexts, memory confidence can play a pivotal role in the investigation of crime and court proceedings. For example, eyewitness confidence is often interpreted by jurors as an indicator of accuracy (Bradfield & Wells, 2000; Brewer & Burke, 2002; Key et al., 2023; Slane & Dodson, 2022). Furthermore, memory confidence is considered influential in determining whether juries believe eyewitness accounts (Cutler et al., 1988, 1990), whereby highly confident eyewitnesses tend to be perceived as more credible (Tenney et al., 2007). However, confident eyewitnesses are not always correct (e.g., Berkowitz et al., 2022) and this can have significant consequences for criminal justice (e.g., Garrett, 2011; Key et al., 2023).

While people can effectively monitor the accuracy of their memories (Koriat & Goldsmith, 1996) by aligning their subjective confidence to objective accuracy, research on eyewitness memory recollection reveals the relationship between confidence and accuracy is fragile. For example, confidence becomes a less reliable predictor of accuracy over time (Goldsmith et al., 2005; Shapira & Pansky, 2019; Spearing & Wade, 2022), after exposure to misinformation (Bohnam & Gonzalez-Vallejo, 2009; Pena et al., 2017; Tomes & Katz, 2000), and when

memory is elicited via deceptive or misleading questions (Howie & Roebers, 2007; see also Brewer et al., 2005). Different types of questions (e.g., free recall and focused questions) can also impact eyewitness confidence and the confidence-accuracy relationship (Allwood et al., 2008). Adults were more confident, more accurate and showed a better confidence-accuracy calibration (i.e., realism) in response to the free recall compared to the focused questions. Similarly, Knutsson et al. (2011) found lowered confidence when probing questions were asked following a free recall prompt.

Therefore, it appears that confidence and its relation to accuracy can be influenced by the type of questions posed during the interview, and the questioning method. This parallels findings from research on eyewitness identification, whereby the relationship between eyewitness confidence and accuracy when making an identification decision is influenced by the line-up procedure. Identifications made with high confidence tend to be associated with high accuracy *only* when optimal conditions are met: such as when confidence judgments are collected immediately (e.g., Brewer & Wells, 2006; Palmer et al., 2013; Sauer et al., 2010), when the identification procedure is unbiased (Wixted & Wells, 2017, see also Wells et al., 2020; Sauer et al., 2019) and in the absence of interviewer feedback (Bradfield et al., 2002; Wells & Bradfield, 1998). While our understanding of eyewitness confidence and the confidence-accuracy relationship in line-up identifications is well advanced, relatively little is known about how various interviewing techniques might influence this relationship in forensic interview contexts.

During a forensic interview, the goal of investigative interviewers is to gather comprehensive and accurate accounts of experienced events using various retrieval-enhancing techniques, including different types of question. Although interviewing techniques should be tailored to eyewitnesses' individual needs, best-practice recommendations suggest initiating an interview with a free recall invitation (Gabbert, Hope, La Rooy, et al., 2015; Gabbert, Hope, McGregor, et al., 2015; Milne & Bull, 1999). This initial prompt provides optimal conditions for accurate retrieval, allowing eyewitnesses to recall the incident in their own words and at their preferred pace. At this stage eyewitnesses tend to be highly accurate (Colomb & Ginet, 2012; Dando et al., 2009; Kontogianni et al., 2020), presumably because they can exert a high level of metacognitive control over their memory output, hence they tend to report details they are confident about, while withholding information they are less sure of (Koriat & Goldsmith, 1996; Weber & Brewer, 2008; see also Allwood et al., 2008; Powell et al., 2005).

Freely reported accounts are typically accurate but do not often include all event-details and so follow-up questions are employed to prompt the retrieval of additional relevant information (Dando, 2013; Dando et al., 2009; Kontogianni et al., 2020; Memon et al., 1997). These questions serve to clarify and expand on aspects of the freely reported account (Fisher, 1995; Fisher & Geiselman, 1992), consequently they typically address event-details mentioned during the free recall phase of the interview. Follow-up questions are described as open-ended breadth and open-ended depth prompts (Powell & Snow, 2007), and can encourage witnesses to provide a more comprehensive narrative. Similarly, open questions starting with the words 'Tell', 'Explain', or 'Describe' (TED questions) (Griffiths & Milne, 2006) encourage eyewitnesses to further expand on elements of their account. Cued-Recall questions—such as those starting with 'Who', 'What', 'Where', 'When', and 'Why', (5-WH questions) (Griffiths & Milne, 2006; Oxburgh et al., 2010) require focused responses, thus they can aid witnesses to search for target-specific details that open questions might have failed to trigger. As such, Cued-Recall questions should also align with the witness's initial free report.

The literature describes this questioning style as *witness-compatible*, and refers to tailoring follow-up questions to the witness's mental representation of the event (Fisher, 2010). Eyewitnesses have unique memories, even for the same event, thus, witness-compatible questions serve to support further recollection simultaneously minimising memory contamination during the interview process. Although the appropriate use of witness-compatible questions can be challenging for interviewers (Fisher, 2010; Fisher & Geiselman, 2010), research shows it is effective in eliciting additional accurate information (e.g., Paulo et al., 2017; also see Maras et al., 2020).

Drawing on this literature, and to better understand eyewitness confidence in relation to accuracy in an interview context, this study investigates the impact of different types of questions, posed in different orders on eyewitness confidence and the confidence-accuracy relationship. Using the mock witness paradigm, participants recalled details of the event with a Free Recall invitation and a set of Cued-Recall questions. Participants in the FR then CR (compatible questions first) received the Free Recall invitation followed by the Cued-Recall questions, further divided into *compatible* questions—focusing on information disclosed in the Free Recall account—and *incompatible* questions, asking for additional information. Similarly, participants in the FR then CR (incompatible questions first), were given a Free Recall followed by the Cued-Recall questions, however, in this group the incompatible questions preceded the compatible questions. Finally, the CR (mixed questions) then FR group received the Cued-Recall questions (randomly presented) followed by the Free Recall invitation. Confidence was measured in two ways: first we collected an aggregate measure of confidence (i.e., confidence in the memory for the entire event seen). In addition, we collected item-by-item confidence judgments in each detail reported in response to the Free Recall invitation and Cued-Recall questions.

While item-by-item confidence allows us to measure the extent to which participants' confidence and accuracy calibrate, the rationale for collecting an aggregate confidence measure is predominantly practical. In real-life, investigators are more likely to inquire about witnesses' confidence in their memory for the entire event as opposed to each individual element, since this is time-consuming and can disrupt the interview. Instead, they might assess confidence after the key phases of the interview. As such, we mirror this approach and collected aggregate confidence across the primary phases of the interview: before the interview (pre-interview confidence) after the Free Recall phase (confidence post phase 1) and after the Cued-Recall phase (post-interview confidence); an additional measure of confidence (confidence post phase 2) was collected after the first set of Cued-Recall questions, for participants in

the FR then CR (compatible questions first) and FR then CR (incompatible questions first) group.

It is reasonable to expect aggregate confidence to be higher after the Free Recall compared to the Cued-Recall phase. This is because in the Free Recall phase participants can exert higher metacognitive control over the retrieval process, and so are more likely to report event-details they are confident about, whereas the Cued-Recall phase is more likely to encourage reporting of details that participants are less sure of. Similarly, we expect compatible questions to lead to (i) higher aggregate confidence and (ii) more effective confidence-accuracy calibration. The rationale for these predictions is that witness-incompatible questions can 'force' participants to verbalise details associated with low confidence, and/or that they cannot remember or may not know. Consequently, it is more likely that confidence for the entire event and the ability to effectively monitor accuracy will be weakened.

2 | METHOD

2.1 | Design

A three (Interview Type: FR then CR (compatible questions first), FR then CR (incompatible questions first) and CR (mixed questions) then FR) x 4 (Confidence Measurement Time: pre-interview confidence, confidence post phase 1, confidence post phase 2, and post-interview confidence) mixed design was used, with Interview Type manipulated between-subjects, and confidence measured within-subjects. The dependent variables for memory reporting were (a) the number of correct details, (b) the number of incorrect details, and (c) the accuracy of the details reported. Accuracy rate was calculated by dividing the number of all details reported by the number of correct details reported. The dependent variables for aggregate confidence were: (d) pre-interview confidence, (e) confidence post phase 1, (f) confidence post phase 2, and (g) post-interview confidence. Finally, the dependent variables for confidence-accuracy calibration were: (h) calibration index (C-index), (j) discrimination (ANDI), and (k) under/overconfidence OU index (for a detailed overview of the calibration analysis see Brewer et al., 2002; Brewer & Wells, 2006).

2.2 | Participants

An a-priori power analysis (G*power, Faul et al., 2007) yielded a sample of 99 participants required to detect a medium effect size $\eta^2_p = .06$ (estimated from Knutsson et al., 2011) with a power (1–β) of .80, and $\alpha = .05$. A large set of data is recommended for the calibration analysis (see Juslin et al., 1996); therefore, based on a previous study whereby 42 participants per condition yielded the recommended size data set, the sample size was increased accordingly. One participant was excluded due to a researcher's mistake in administering the study procedure, leaving a total of 124 participants in the study (females = 100, Mean age = 21.85, SD = 5.05). Participants were recruited among university students and administrative staff.

3 | MATERIALS

3.1 | Video stimulus

The mock crime video (1 min, 40 s) was presented on a 17-inch HD screen. The event depicted a non-violent robbery filmed in a store and involved four males. A customer is seen entering the shop and asking the shopkeeper for directions. Shortly after, two Caucasian males enter and proceed straight into a hidden section where they disguise their identity and then proceed to rob the shop. After a few seconds, a customer intervenes and one of the robbers pushes him to the floor. Finally, the robbers take their disguises off and run away.

3.2 | The interview phases

Three interviews were used: (i) the FR then CR (compatible questions first), (ii) the FR then CR (incompatible questions first) and, (iii) the CR (mixed questions) then FR. All interviews encompassed two phases: the Free Recall and Cued-Recall phase.

Free Recall phase. Here, participants were asked to imagine the researcher knew nothing about the crime event they had seen and to write down all the details they could remember. To allow participants' full control over their memory output, no further instruction was given at this stage.

Cued-Recall phase. This phase was administered verbally and consisted of 89 predetermined Cued-Recall questions relating to different aspects of the crime, including the people involved, their actions, their descriptions, and any personal effects they had with them. Participants were asked to always answer the question even when unsure.

3.3 | The order of the interview phases

The order of the interview phases was different depending on condition. In the FR then CR (compatible questions first) interview, the Free Recall phase preceded the Cued-Recall phase. In the Cued-Recall phase, the 89 Cued-Recall questions were administered as follows: the *compatible* questions preceded the *incompatible* questions. A question was considered *compatible* when related to a detail reported by the participant in the previous Free Recall account. For example, if in the Free Recall the participant mentioned the 'dark-haired robber', all the Cued-Recall questions related to the dark-haired robber's description were considered compatible (e.g., What was the dark-haired robber ethnicity? How old was the dark-haired robber?). Similarly, if in the Free Recall, they mentioned the robber's 'disguise', all the Cued-Recall questions related to the robber's disguise were considered compatible (e.g., What type of disguise was the robber wearing?). Once the compatible questions were selected all remaining questions of the pool were considered *incompatible*.

In the FR then CR (incompatible questions first) interview, the Free Recall phase preceded the Cued-Recall phase. In the Cued-Recall

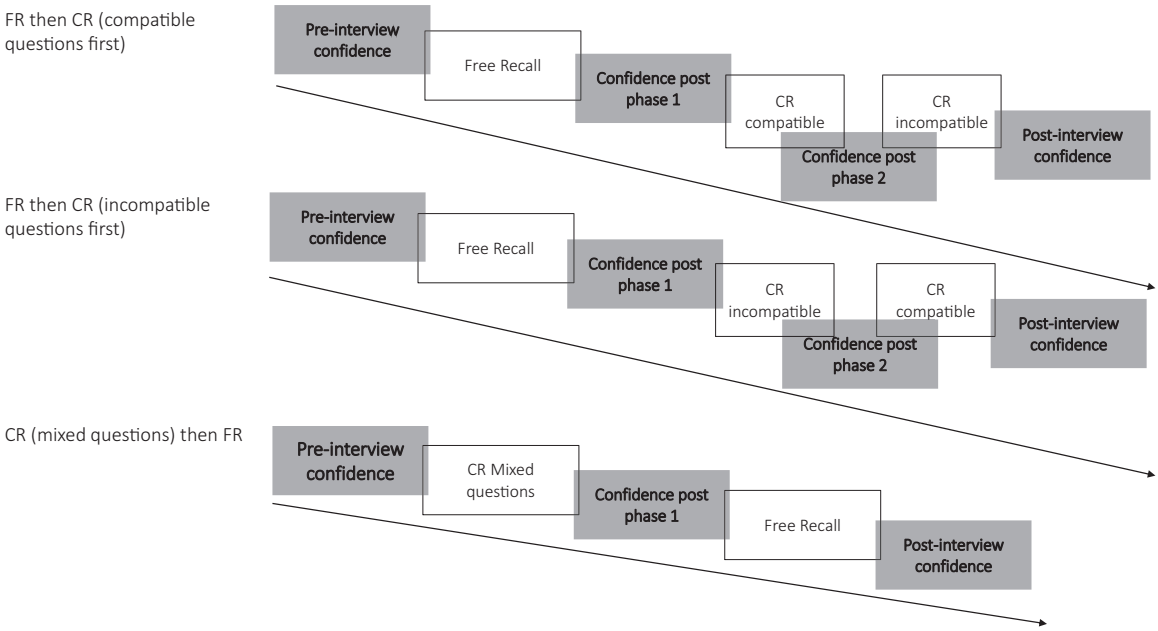

FIGURE 1 Study procedure.

phase participants received the incompatible questions followed by the compatible questions.[1]

In the CR (mixed questions) then FR, the Cued-Recall phase preceded the Free Recall phase. During the Cued-Recall phase, the 89 Cued-Recall questions were randomly presented.

Regardless of experimental condition, participants in all interview groups received a Free Recall invitation and the 89 predetermined Cued-Recall questions, albeit in different orders.

3.4 | Confidence

Aggregate confidence in the accuracy of the memory for the entire event was measured (i) before the interview (i.e., pre-interview confidence), (ii) after the initial phase of the interview (i.e., confidence post phase 1), (iii) after the complete interview (post-interview confidence). An additional aggregate confidence judgment (i.e., confidence post phase 2) was collected after the first set of Cued-Recall questions (see Figure 1) for the FR then CR (compatible questions first) and the FR then CR (incompatible questions first) group. In addition, item-by-item confidence was collected after each response to the 89 Cued-Recall questions, and for each detail reported in the Free Recall phase. All confidence ratings were collected on a scale ranging from 0% to 100%, whereby 0% represented 'not at all confident', and 100% represented 'completely confident' in the answer reported.

3.5 | Procedure

Participants were tested individually in a research laboratory on the university campus. After being presented with an initial debrief detailing the procedure of the study, and a consent form, participants watched the crime video, immediately followed by a three-minute filler task. From here on the procedure differed for each group.

Participants in the FR then CR (compatible questions first), were presented with the Free Recall invitation followed by a 7-min filler task. While the participants completed the filler task the researcher highlighted each detail reported in the Free Recall account. For example, if the participant reported 'The robber put the money in a plastic bag', the research highlighted the details: (1) robber, (2) put the money, (3) plastic, and (4) bag. After the filler task, participants provided a confidence rating on each detail highlighted in their Free Recall account. Simultaneously, two research assistants identified the compatible and incompatible questions out of the pool of 89 Cued-Recall questions. Once participants had provided a confidence rating on the details of their Free Report, they were given the 89 Cued-Recall questions in this order: the compatible questions were asked *before* the incompatible questions. In this group aggregate confidence was measured on four occasions: before the interview (pre-interview confidence), after the Free Recall phase (confidence post phase 1), after the compatible questions (confidence post phase 2), and after the incompatible questions (post-interview confidence).

The procedure for the FR then CR (incompatible questions first) was identical, but the 89 Cued-Recall questions were presented in the opposite order, that is the incompatible questions preceded the compatible question. Aggregate confidence in this group was measured on four occasions: before the interview (pre-interview confidence) after the Free Recall phase (confidence post phase 1), after the incompatible (confidence post phase 2), and after the compatible questions (post-interview confidence).

Participants in the CR (mixed questions) then FR group were given the 89 Cued-Recall questions presented in random order,

followed by the Free Recall invitation. After the Free Recall, participants completed the seven-minute filler task, to allow time for the researcher to highlight all event-details of the Free Recall account. After the filler task, participants provide a confidence rating on each detail highlighted in their Free Recall account. In this group, aggregate confidence was measured on three occasions: before the interview (pre-interview confidence), after the 89 Cued-Recall questions (confidence post phase 1), and after the Free Recall phase (confidence after phase 2). To control for interviewer variability, all interviews were conducted by the same researcher (the principal investigator) using the condition-appropriate protocol (verbatim—detailed interview protocols are available from the first author). Prior data collection the experimental procedure was piloted (this data was not included in the final dataset). The experiment took between 60 and 75 min per participant.

3.6 | Coding

Each detail reported in response to the Free Recall request was coded as 'correct' if present in the video, or as 'incorrect' if not present in the video. Non-specific details (e.g., 'he said something') and personal opinions (e.g., 'I think they were accomplices') were not coded. An independent researcher coded 20% accounts, the Interclass Correlation Coefficients were as follow: correct (ICC = 0.92, p < 001), incorrect (ICC = 0.94, p < 001), and accuracy of the details reported (ICC = 0.93, p < .001).

4 | RESULTS

4.1 | Stability of aggregate confidence across the interview phases

First, we investigated potential shifts in aggregate confidence across the interview phases. We performed a 3 [Interview Type: FR then CR (compatible questions first), FR then CR (incompatible questions first) and CR (mixed questions) then FR] x 3 (Confidence Measurement Time: pre-interview confidence, confidence post phase 1, and post-interview confidence) mixed ANOVA, with confidence measured within-subjects. There was a significant main effect of Measurement Time on confidence $F(1.87, 226.83) = 22.83$, p < .001, $\eta^2_p = .16$, a significant main effect of Interview Type $F(2, 121) = 4.05$, $p = .02$, $\eta^2_p = .06$, and a significant interaction $F(3.75, 226.83) = 11.51$, p < .001, $\eta^2_p = .16$ (see Table 1 for Means and SD). Bonferroni post hoc comparisons showed that pre-interview confidence did not differ significantly between groups. However, after phase 1, confidence reported by participants in the CR (mixed questions) then FR group was significantly lower than that reported by participants in FR then CR (compatible questions first) ($M_{diff} = -22.48$, p < .001, 95% CI [−32.79, −12.18]), and FR then CR (incompatible questions first) group ($M_{diff} = -18.29$, p < .001, 95% CI [−28.66, −7.92]). No other significant difference was found at this stage.

After phase 1, only participants in the CR (mixed questions) then FR group reported significantly lower confidence compared to their pre-interview confidence ($M_{diff} = -17.31$, p < .001, 95% CI [−24.46, −10.17]). On the contrary, participants in the remaining two groups did not report significant changes between pre-interview confidence and confidence post phase 1.

No difference between groups was found on post-interview confidence. After the interview, participants in the FR then CR (compatible questions first), and FR then CR (incompatible questions first) group reported significantly decreased confidence compared to confidence post phase 1 (respectively $M_{diff} = -13.33$, p < .001, 95% CI [−19.88, −6.78], and $M_{diff} = -16.58$, p < .001, 95% CI [−23.21, −9.95]); on the contrary participants in the CR (mixed questions) then FR group reported significantly increased confidence compared to confidence post phase 1 ($M_{diff} = 8.53$, $p = .007$, 95% CI [1.91, 15.16]).

In summary and as predicted, aggregate confidence was higher after the Free Recall (cf. Cued-Recall) phase. In addition, when the interview featured an initial Free Recall invitation followed by a Cued-Recall phase, aggregate confidence remained stable after the former phase and was more likely to decrease after the latter.

4.2 | Stability of aggregate confidence in the Cued-Recall phase

Next, we investigated potential shifts in confidence after the set of compatible and incompatible questions for the FR then CR (compatible questions first) and FR then CR (incompatible questions first) groups only. A 2 [Interview Type: FR then CR (compatible questions first), and FR then CR (incompatible questions first)] × 4 (Confidence Measurement Time: pre-interview confidence, confidence post phase

TABLE 1 Means (and SD) of pre-interview confidence, confidence post phase 1, confidence post phase 2 (measured only in for the FR then CR (compatible questions first) and the FR then CR (incompatible questions first)), and post-interview confidence for participants in all groups.

	Cr (mixed questions) then FR	FR then CR (compatible questions first)	FR then CR (incompatible questions first)
Pre-interview confidence	62.93 (16.77)	66.67 (14.08)	61.71 (14.98)
Confidence post phase 1	45.61 (20.50)	68.10 (18.38)	63.90 (19.08)
Confidence post phase 2	-	68.57 (13.53)	42.46 (18.95)
Post-interview confidence	54.15 (21.56)	54.76 (19.78)	47.32 (18.31)

1, confidence post phase 2, and post-interview confidence) mixed ANOVA was conducted. We found a significant main effect of Interview Type, $F(1, 81) = 13.21$, $p < .001$, $\eta^2_p = .14$, a significant main effect of Measurement Time on confidence $F(2.50, 203.09) = 26.22$, $p < .001$, $\eta^2_p = .25$, and a significant interaction $F(2.50, 203.09) = 14.03$, $p < .001$, $\eta^2_p = .15$. Bonferroni post hoc comparisons showed no differences between the two interview groups on confidence reported at the pre-interview and post phase 1 stage. However, a significant difference was found on confidence reported at the post phase 2 stage; here confidence for participants answering the incompatible questions was significantly lower than confidence for those answering the compatible questions ($M_{diff} = -26.13$, $p < .001$, 95% CI [−33.31, −18.96]).

At the post phase 2 stage, confidence for participants answering incompatible questions decreased significantly compared to the previous phase 1 ($M_{diff} = -21.46$, $p < .001$, 95% CI [−29.66, −13.26]), while confidence for participants answering compatible questions did not change significantly. Finally, no difference between groups was found on confidence reported at a post-interview stage. At a post-interview stage, participants answering the set of incompatible questions reported significantly lowered confidence compared to the previous post phase 2 stage ($M_{diff} = -13.81$, $p < .001$, 95% CI [−19.89, −7.72]), while those answering the set of compatible questions increased their confidence, but importantly, this increase was not significant ($M_{diff} = 4.87$, $p = .21$, 95% CI [−1.28, 11.03]).

In summary, as we anticipated participants reported lower aggregate confidence after answering the set of incompatible (cf. compatible) questions. In addition, within-subject comparisons showed that when the Cued-Recall phase featured compatible followed by incompatible questions, confidence remained stable only until the incompatible questions were posed.

4.3 | Confidence for compatible and incompatible questions

Next, we examined item-by-item confidence reported in response to the compatible and incompatible questions. Previous research shows that repeatedly reporting a memory can inflate confidence (i.e., a reiteration effect, see Knutsson et al., 2011). To control for this phenomenon, we performed the following analyses *only* on the new event-details reported in the Cued-Recall phase; that is, we excluded all the event-details reported in response to the Cued-Recall questions if these were also reported in the Free Recall phase.

A 2 (compatibility of questions: compatible vs. incompatible) × 2 (order of questions: FR then CR [compatible questions first], and FR then CR [incompatible questions first]) mixed ANOVA, with compatibility of questions as a within-subject variable and order of questions as a between-subject variable, showed a significant main effect of compatibility on confidence $F(1, 81) = 229.92$, $p < .001$, $\eta^2_p = .74$, meaning that independently on the order in which the questions were asked, confidence for details reported in response to compatible questions was higher than confidence for details reported in response to incompatible questions ($M_{diff} = 23.81$, $p < .001$, 95% CI [20.74, 27.01]). We also found a marginally significant main effect of order of question $F(1, 81) = 4.35$, $p = .04$ $\eta^2_p = .05$; meaning that confidence across all Cued-Recall questions was higher in the FR then CR (compatible questions first) (cf. FR then CR [incompatible questions first]) ($M_{diff} = 6.33$, $p < .04$, 95% CI [.29, 12.38])—that is when the compatible questions preceded the incompatible questions. Finally, we found a non-significant interaction $F(1, 81) = .14$, $p = .71$.

Following this, a calibration analyses was performed to examine how the compatibility and order of questions affected the reliability of confidence judgements. A series of 2 (compatibility of questions: compatible vs. incompatible) × 2 [order of questions: FR then CR (compatible questions first) vs. FR then CR (incompatible questions first)] mixed design ANOVAs were performed on (a) C-index, (b) ANDI, and (c) OU measures. C-index is a measure of calibration, it ranges from 0 to 1, where 0 represents perfect calibration. On the C-index we found a significant main effect of compatibility of questions $F(1, 81) = 5.53$, $p = .02$, $\eta^2_p = .06$. The main effect of order of question was non-significant $F(1, 81) = .6$, $p = .44$; however, we found a significant interaction $F(1, 81) = 4.98$, $p = .03$, $\eta^2_p = .06$. Bonferroni follow-up comparisons showed that the C-index for compatible questions was lower than that for incompatible questions ($M_{diff} = -.05$, $p = .002$, 95% CI [−.08, −.02]) in the FR then CR (compatible questions first) group *only*; meaning that calibration on the compatible questions was better than calibration for incompatible questions *only* when the former questions preceded the latter (see Table 2 for Means and SD).

ANDI is a measure of discrimination, it ranges from 0 to 1, where 1 represents perfect discrimination. For this measure, we found a significant main effect of compatibility of questions $F(1, 81) = 26.14$, $p < .001$, $\eta^2_p = .24$, meaning that independently of the order in which the questions were asked, discrimination was higher for compatible compared to incompatible questions ($M_{diff} = .19$, $p < .001$, 95% CI [.12, .27]). The main effect of order was not significant $F(1, 81) = .72$, $p = .4$, nor was the interaction $F(1, 81) = .67$, $p = .42$.

Finally, the OU index summaries the degree of under/overconfidence, this value ranges from – 1 to +1, where negative values represent under-confidence and positive values signify overconfidence. Here, we found a non-significant main effect of compatibility $F(1, 81) = .50$, $p = .48$, and a significant main effect of order of questions $F(1, 81) = 5.13$, $p = .03$, $\eta^2_p = .06$. This latter result shows that across all the Cued-Recall questions (compatible and incompatible) participants in the FR then CR (compatible questions first) group were less underconfident than those in the FR then (incompatible questions first) group ($M_{diff} = .07$, $p = .03$, 95% CI [0.01, 0.14]). Finally, the interaction was non-significant $F(1, 81) = 0.03$, $p = 0.87$.

In summary, confidence in new event-details reported in the Cued-Recall phase was higher when these were extracted via compatible (cf. incompatible) questions. More importantly, compatible questions promoted better calibration (lower C-index, less negative OU index) and more effective discrimination (higher ANDI). The order in which the questions were presented affected calibration but not discrimination. Calibration was more effective in response to compatible

TABLE 2 Means (and SD) of C-index, ANDI, and OU measures for event-details reported in response to compatible and incompatible questions in the two groups.

		FR then CR (compatible questions first)	FR then CR (incompatible questions first)
C-index	Compatible	0.08 (0.05)	0.12 (0.09)
	Incompatible	0.13 (0.08)	0.12 (0.05)
OU measure	Compatible	−0.03 (0.13)	−0.10 (19)
	Incompatible	−0.01 (0.19)	−0.09 (0.19)
ANDI	Compatible	0.43 (0.26)	0.37 (0.26)
	Incompatible	0.21 (0.26)	0.21 (0.18)

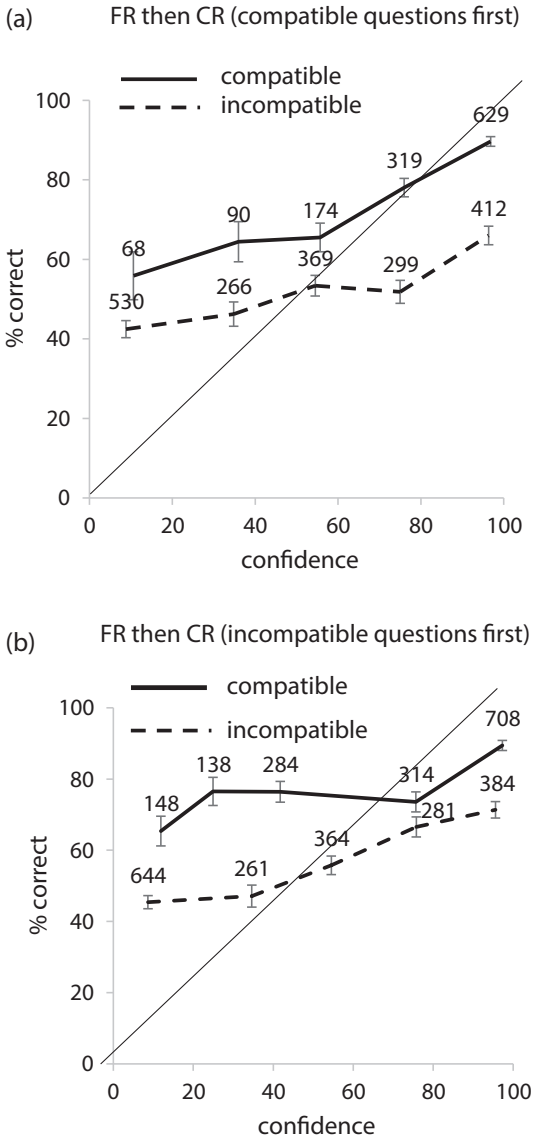

FIGURE 2 Calibration curves (and data points for each confidence interval) for the compatible and incompatible questions for participants in the FR then CR (compatible questions first) (Panel a), and FR then CR (incompatible questions first) group (Panel b).

questions when these preceded the incompatible questions. This pattern of results can be observed in the calibration curves (Figure 2). The solid line—representing calibration for the compatible questions, is closer to the diagonal line (representing perfect calibration) in the FR then CR (compatible questions first) group (Panel a). Furthermore, only for the FR then CR (compatible questions first) group is the solid line closer to the diagonal line than the dashed line—representing calibration on the incompatible questions.

4.4 | Memory report

While memory reporting was not the primary focus of this study, to report a complete picture, we analysed the impact of compatible and incompatible questions on memory output. There was no significant difference between groups in the number of correct details $t(81) = 1.94$, $p = .06$, number of incorrect details $t(81) = 1.58$, $p = .12$, nor the accuracy $t(81) = -.83$, $p = .41$ of the details reported in the Free Recall phase. Following this, the impact of compatibility and order of questions was analysed with a 2 (compatibility of questions: compatible vs. incompatible) × 2 [order of questions: FR then CR (compatible questions first), and FR then CR (incompatible questions first)] mixed design ANOVA.

On the number of correct details we found a significant main effect of compatibility of questions $F(1, 81) = 31.34$, $p < .001$, $\eta^2_p = .28$, a non-significant main effect of order $F(1, 81) = .26$, $p = .61$, and a significant interaction $F(1, 81) = 5.63$, $p = .02$, $\eta_p^2 = .07$ (see Table 3 for Means and SD). Bonferroni follow-up comparisons showed that participants in the FR then CR (compatible questions first) group reported significantly more correct details in answer to compatible (cf. incompatible) questions ($M_{diff} = 12.00$, $p < .001$, 95% CI [7.79, 16.21]), as did those in the FR then CR (incompatible questions first) group ($M_{diff} = 4.85$, $p = .02$, 95% CI [5.93, 9.11]). However, when the same analysis was conducted *only* on new correct details reported, these results were statistically non-significant.

On number of incorrect details, we found a significant main effect of compatibility of questions $F(1, 81) = 231.16$, $p < .001$, $\eta^2_p = .74$; here, independently of the order in which the questions were presented, compatible questions elicited fewer incorrect details compared to the incompatible questions ($M_{diff} = -14.15$, $p < .001$, 95% CI [−16.00, −12.29]). The main effect of order was non-significant $F(1, 81) = .19$, $p = .66$, nor was the interaction $F(1, 81) = .29$, $p = .59$. This pattern of results was similar when only new incorrect details were included in the analysis.

TABLE 3 Means (and SD) of number of correct, incorrect details, and accuracy rate of the details reported in response to the Free Recall and the compatible and incompatible questions in both groups.

		FR then CR (compatible question first)	FR then CR (incompatible questions first)
Free Recall	Correct	47.52 (14.92)	41.29 (14.27)
	Incorrect	3.79 (3.28)	2.78 (2.46)
	Accuracy rate	92.74 (5.43)	93.73 (5.52)
Compatible	Correct	35.83 (8.90)	31.85 (7.49)
	Incorrect	7.21 (3.48)	7.05 (4.19)
	Accuracy rate	83.31 (7.39)	82.31 (8.89)
Incompatible	Correct	23.83 (6.64)	27.00 (7.77)
	Incorrect	20.86 (7.31)	21.72 (6.29)
	Accuracy rate	53.71 (11.20)	55.15 (10.85)

Finally, on the accuracy rate we found a significant main effect of compatibility of questions $F(1,81) = 440.96$, $p < .001$, $\eta^2_p = .85$, meaning that regardless of the order in which the questions were presented, participants were significantly more accurate in response to compatible than in response to incompatible questions ($M_{diff} = 28.29$, $p < .001$, 95% CI [25.61, 30.98]). The main effect of order was non-significant $F(1, 81) = .03$ $p = .86$, nor was the interaction $F(1, 81) = .72$, $p = .39$. This pattern of results remained unchanged when only new event-details were included in the analysis.

In summary, compatible (cf. incompatible) questions yielded more correct event-details, fewer errors, and higher accuracy rate overall. The order in which the questions were asked influenced the number of correct details reported overall. Thus, initial evidence suggests that compatible questions might elicit more correct event-details only when they precede the incompatible questions. The data set of this research study is available on the OSF (https://osf.io/xvrfq/?view_only=bd96232f94984ffb893daec44b9c1fe2).

5 | DISCUSSION

This study investigated the impact of question type and question order on eyewitness memory confidence and its relationship to memory accuracy. Here, a Free Recall invitation and compatible Cued-Recall questions promoted higher and more stable confidence for the entire event, compared to incompatible questions. Compatible questions were more likely to elicit (i) accurate additional information, (ii) fewer errors, and (iii) more reliable confidence-accuracy calibration and discrimination, especially when they preceded incompatible questions. Together, our findings offer strong support for encouraging interviewers to adhere to best practice of eliciting a Free Recall account followed by relevant probing questions to expand on topics mentioned in the Free Recall. Considering the low accuracy, and poor confidence-accuracy calibration associated with incompatible questions, results indicate it is advisable to limit the use of incompatible questions.

Our findings support the prediction that confidence in the memory for the entire event would be higher after the Free Recall phase compared to the Cued-Recall phase. Regardless of whether the Free Recall phase preceded the Cued-Recall phase or not, participants reported higher confidence in their memory following a Free Recall invitation. Confidence decreased only when the Cued-Recall questions did not relate to elements reported spontaneously by the eyewitness. Conversely, when the Cued-Recall questions were compatible, focusing on event-details mentioned in the Free Recall, global confidence was more likely to remain stable. Overall, our results suggest that where interviews (i) follow the recommended UK guidance (Gabbert, Hope, La Rooy, et al., 2015; Gabbert, Hope, McGregor, et al., 2015; Home Office, 2022; Milne & Bull, 1999) whereby the Free Recall invitation precedes the Cued-Recall phase, and (ii) a witness-compatible questioning style is adopted, confidence is more likely to remain stable. On the contrary, when an interview features incompatible questions, global confidence decreases. These results can be explained by considering the nature of the incompatible questions and the event-details they tend to trigger, and the cues upon which aggregate confidence is constructed.

Compared to compatible questions and Free Recall invitations, incompatible questions tend to direct attention to elements of the event that could be unknown, currently inaccessible, or associated with low confidence; hence, these questions may be more likely to expose potential gaps in memory which in turn could affect overall confidence. Research on metamemory distinguishes between aggregate judgments (i.e., post-test performance estimate or PTPE), which are global estimates of the items of a test the participant believed to have recalled, and item-by-item judgments, that refer to the recollection of one item out of a list of items presented (Mazzoni & Nelson, 1995; Schneider et al., 2000; see also Fu et al., 2012). The literature suggests that these two types of judgment are likely constructed on different cues (Fu et al., 2012). When making item-by-item judgments, participants tend to evaluate cues relating to the recollection of the single item in hand (Koriat et al., 2008; see also Robinson et al., 2000). In contrast, when constructing aggregate judgments, participants might adopt a broader perspective, which may include considering previous attempts to recall similar items. Therefore, it is possible that when participants formed global confidence judgments, they considered among other factors their most recent

performance in recalling the event. Considering performance was far superior in response to the Free Recall and the compatible questions compared to the incompatible questions (average accuracy was 92.64% for the Free Recall; 82.74% for compatible, and 54.43% for the incompatible questions), it is reasonable to expect that only the latter type of question promoted the impression of poorer memory. This might also explain why confidence reported by participants in the FR then CR (incompatible questions first) group did not rise significantly following compatible questions. It is possible that at this stage global confidence encompassed not only performance on the compatible questions but also their lower performance on the preceding incompatible questions.

In real-life, it may not be feasible for investigators to inquire about witness confidence following each event-detail reported, as this can interrupt episodic retrieval. However, investigators might ask for confidence indicators after the key phases of the interview, or once they had the opportunity to ask questions and gain an initial understanding of the event or important elements, such as the description of a perpetrator. Therefore, it is important to consider whether global judgments encompass overall recollection and so future research should elucidate if confidence dented by an initial interview might persist over time. For example, in time-critical situations when investigators must gather specific details as quickly as possible and so might ask focused questions; might this lower global confidence, and if so does low confidence persist over time? Additionally, it is feasible that diminishing confidence might affect witnesses' attitudes towards accepting additional interview requests or engaging further with the Justice System, such as testifying in Court. Lowered confidence might reduce witnesses' self-efficacy within the legal system, with potential consequences for their well-being (Diesen, 2012) and suggestibility (Jaeger et al., 2012; Leippe et al., 2006).

Our results indicate compatible questions tend to elicit (i) more accurate event-details, (ii) fewer errors, and (iii) more reliable confidence-accuracy calibration and discrimination. These findings are in line with previous research showing that probing questions asked after an initial Free Recall invitation, can successfully extract more comprehensive accounts (Dando et al., 2009; Kontogianni et al., 2020). Similarly, here compatible questions elicited, on average, 35.58% of the correct event-details reported during the interview. However, we present initial evidence that compatible questions, if not preceded by incompatible questions, may also promote a more reliable confidence-accuracy calibration, especially when participants express high confidence (80%–100%). Existing literature shows that witnesses' confidence and accuracy are better calibrated for event-details they are reasonably confident about (Allwood et al., 2008), and that event-details associated with high confidence tend to be reported in the Free Recall account (Saraiva et al., 2020); it is therefore likely that questions focusing exclusively on information freely reported also promote a more effective alignment between confidence and accuracy.

These findings indicate that the memory confidence and accuracy alignment might depend, in part, upon the way a memory is cued during an interview. The structure of the interview phases, the sequential organisation of the Cued-Recall questions, and their compatibility with the free account might all contribute towards shaping confidence and its relation to accuracy. However, caution is warranted because even when the Free Recall phase preceded the Cued-Recall phase and the latter phase featured compatible questions, the relationship between confidence and accuracy was far from perfect. An observation of the curve for compatible questions in the FR then CR (compatible questions first) group shows both a slight overconfidence associated with very high confidence (80%–100%) and a noticeable under-confidence associated with moderate to low confidence (50%–0%). It is widely acknowledged that overconfidence can lead to erroneous convictions (Berkowitz et al., 2022), and under-confidence can be equally problematic. For example, underconfident eyewitnesses may hesitate to share accurate information, or they might report information but verbalise their uncertainty which might deter investigators from pursuing crucial leads. Since limited research has explored the relationship between confidence and accuracy in the interviewing context, practical recommendations are premature.

There are some methodological limitations common to research of this nature. For instance, our participants were presented with a large set of Cued-Recall questions. While this was necessary to collect sufficient data for our calibration analysis, participant concentration may not have remained stable throughout the interview. In real-life, interviewers typically employ a combination of questions, including open prompts, or they might use compatible and incompatible questions in varied sequences, unlike the controlled order used in our study. Future research could address these limitations and their potential impact on confidence and its relationship to accuracy. Nevertheless, our results provide an important initial understanding of the impact that different types of question posed in different orders can have on eyewitness confidence and shed further light on how investigative interviewers might promote a more effective alignment between memory confidence and accuracy.

CONFLICT OF INTEREST STATEMENT

The authors have no conflicts of interest to disclose.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in OSF at https://osf.io/xvrfq/?view_only=bd96232f94984ffb893daec44b9c1fe2.

ORCID

Alessandra Caso https://orcid.org/0000-0002-7731-3860
Fiona Gabbert https://orcid.org/0000-0002-2727-1113

ENDNOTE

[1] Participants in the FR then CR (compatible questions first) and FR then CR (incompatible questions first) answered on average 41 compatible questions (average SD = 9.42; range was respectively 41 and 40), there was no difference in the number of compatible questions answered between groups (M_{diff} = 4.14, $t(81)$ = 2.01, p = .05, 95% CI [0.26, 8.26]).

REFERENCES

Allwood, C. M., Innes-Ker, A. C., Homgren, J., & Fredin, G. (2008). Children's and adults' realism in their event-recall confidence in responses

to free recall and focused questions. *Psychology, Crime & Law*, *14*(6), 529–547. https://doi.org/10.1080/10683160801961231

Berkowitz, S. R., Garrett, B. L., Fenn, K. M., & Loftus, E. F. (2022). Convicting with confidence? Why we should not over-rely on eyewitness confidence. *Memory*, *30*(1), 10–15. https://doi.org/10.1080/09658211.2020.1849308

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

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

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

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

Brewer, N., Keast, A., & Rishworth, A. (2002). The confidence-accuracy relationship in eyewitness identification: The effects of reflection and disconfirmation on correlation and calibration. *Journal of Experimental Psychology: Applied*, *8*(1), 44–56. https://doi.org/10.1037/1076-898X.8.1.44

Brewer, N., & Wells, G. L. (2006). The confidence-accuracy relationship in eyewitness identification: Effects of lineup instructions, foil similarity, and target-absent base rates. *Journal of Experimental Psychology: Applied*, *12*(1), 11–30. https://doi.org/10.1037/1076-898X.12.1.11

Brewer, W. F., Sampaio, C., & Barlow, M. R. (2005). Confidence and accuracy in the recall of deceptive and nondeceptive sentences. *Journal of Memory and Language*, *52*(4), 618–627. https://doi.org/10.1016/j.jml.2005.01.017

Colomb, C., & Ginet, M. (2012). The cognitive interview for use with adults: An empirical test of an alternative mnemonic and of a partial protocol. *Applied Cognitive Psychology*, *26*(1), 35–47. https://doi.org/10.1002/acp.1792

Cutler, B. L., Dexter, H. R., & Penrod, S. D. (1990). Non adversarial methods for sensitizing jurors to eyewitness evidence. *Journal of Applied Social Psychology*, *20*(14), 1197–1207. https://doi.org/10.1111/j.1559-1816.1990.tb00400.x

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

Dando, C. (2013). Drawing to remember: External support of older adults' eyewitness performance. *PLoS One*, *8*, e69937. https://doi.org/10.1371/journal.pone.0069937

Dando, C., Wilcock, R., Milne, R., & Henry, L. (2009). An adapted cognitive interview procedure for frontline police investigators. *Applied Cognitive Psychology*, *23*, 698–716. https://doi.org/10.1002/acp.1501

Diesen, C. (2012). Therapeutic jurisprudence and the victim of crime. In T. I. Oei & M. S. Groenhuijsen (Eds.), *Progression in forensic psychiatry: About boundaries* (pp. 579–598). Kluwer.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*power 3: A flexible statistical power analysis for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/bf03193146

Fisher, R. P. (1995). Interviewing victims and witnesses of crime. *Psychology, Public Policy, and Law*, *1*(4), 732–764. https://doi.org/10.1037/1076-8971.1.4.732

Fisher, R. P. (2010). Interviewing cooperative witnesses. *Legal and Criminological Psychology*, *15*, 25–38. https://doi.org/10.1348/135532509X441891

Fisher, R. P., & Geiselman, R. E. (1992). *Memory-enhancing techniques for investigative interviewing: The cognitive interview*. Charles C Thomas, Publisher.

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

Fu, T., Koutstaal, W., Poon, L., & Cleare, A. J. (2012). Confidence judgment in depression and dysphoria: The depressive realism vs. negativity hypotheses. *Journal of Behavior Therapy and Experimental Psychiatry*, *43*(2), 699–704. https://doi.org/10.1016/j.jbtep.2011.09.014

Gabbert, F., Hope, L., La Rooy, D. J., McGregor, A., Milne, R., & Ellis, T. (2015). Introducing a PEACE-compliant 'structured interview protocol' to enhance the quality of investigative interviews. iIIRG. https://www.iiirg.org/assets/Gabbert-et-al.-SIP-talk-for-IIRG-pdf

Gabbert, F., Hope, L., McGregor, A., Ellis, T., La Rooy, D. J., & Milne, R. (2015). SIP Structured Interview Protocol. [Pamphlet].

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

Goldsmith, M., Koriat, A., & Pansky, A. (2005). Strategic regulation of grain size in memory reporting over time. *Journal of Memory and Language*, *52*(4), 505–525. https://doi.org/10.1016/j.jml.2005.01.010

Griffiths, A., & Milne, R. (2006). Will it all end in tiers? Police interviews with suspects in Britain. In T. A. Williamson (Ed.), *Investigative interviewing: Rights, research, regulation* (pp. 167–189). Willan.

Home Office. (2022). Achieving best evidence in criminal proceedings: Guidance on interviewing victims and witnesses, and guidance on using special measures.

Howie, P., & Roebers, C. M. (2007). Developmental progression in the confidence-accuracy relationship in event recall: Insights provided by a calibration perspective. *Applied Cognitive Psychology*, *21*(7), 871–893. https://doi.org/10.1002/acp.1302

Jaeger, A., Lauris, P., Selmeczy, D., & Dobbins, I. G. (2012). The costs and benefits of memory conformity. *Memory & Cognition*, *40*(1), 101–112. https://doi.org/10.3758/s13421-011-0130-z

Juslin, P., Olsson, N., & Winman, A. (1996). Calibration and diagnosticity of confidence in eyewitness identification: Comments on what can be inferred from the low confidence–accuracy correlation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(5), 1304–1316. https://doi.org/10.1037/0278-7393.22.5.1304

Key, J. S., Neuschatz, S. D., Gronlund, D. D., Wetmore, S. A., McAdoo, R. M., & McCollum, D. (2023). High eyewitness confidence is always compelling: that's a problem. *Psychology, Crime & Law*, *29*(1), 120–141. https://doi.org/10.1080/1068316X.2021.2007912

Knutsson, J., Allwood, C. M., & Johansson, M. (2011). Child and adult witnesses: The effect of repetition and invitation-probes on free recall and metamemory realism. *Metacognition and Learning*, *6*(3), 213–228. https://doi.org/10.1007/s11409-011-9071-y

Kontogianni, F., Hope, L., Taylor, P. J., Vrij, A., & Gabbert, F. (2020). "Tell me more about this…": An examination of the efficacy of follow-up open questions following an initial account. *Applied Cognitive Psychology*, *34*(5), 972–983. https://doi.org/10.1002/acp.3675

Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review*, *103*(3), 490–517. https://doi.org/10.1146/annurev.psych.51.1.481

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

Leippe, M. R., Eisenstadt, D., Rauch, S. M., & Stambush, M. A. (2006). Effects of social-comparative memory feedback on eyewitnesses' identification confidence, suggestibility, and retrospective memory reports. *Basic and Applied Social Psychology*, *28*(3), 201–220. https://doi.org/10.1207/s15324834basp2803_1z

Maras, K., Dando, C., Stephenson, H., Lambrechts, A., Anns, S., & Gaigg, S. (2020). The witness-aimed first account (WAFA): A new technique for interviewing autistic witnesses and victims. *Autism*, *24*(6), 1449–1467. https://doi.org/10.1177/1362361320908986

Mazzoni, G., & Nelson, T. O. (1995). Judgments of learning are affected by the kind of encoding in ways that cannot be attributed to the level of recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(5), 1263–1274. https://doi.org/10.1037/0278-7393.21.5.1263

Memon, A., Wark, L., Holley, A., Bull, R., & Koehnken, G. (1997). Eyewitness performance in cognitive and structured interviews. *Memory*, *5*(5), 639–656. https://doi.org/10.1080/741941481

Milne, R., & Bull, R. (1999). *Investigative interviewing: Psychology and practice*. Wiley.

Oxburgh, G. E., Myklebust, T., & Grant, T. D. (2010). The question of question types in police interviews: A review of the literature from a psychological and linguistic perspective. *International Journal of Speech, Language and the Law*, *17*(1), 45–66. https://doi.org/10.1558/ijsll.v17i1.45

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

Paulo, R. M., Albuquerque, P. B., Vitorino, F., & Bull, R. (2017). Enhancing the cognitive interview with an alternative procedure to witness-compatible questioning: Category clustering recall. *Psychology, Crime & Law*, *23*(10), 967–982. https://doi.org/10.1080/1068316X.2017.1351966

Pena, M. M., Klemfuss, J. Z., Loftus, E. F., & Mindthoff, A. (2017). The effects of exposure to differing amounts of misinformation and source credibility perception on source monitoring and memory accuracy. *Psychology of Consciousness: Theory, Research and Practice*, *4*(4), 337–347. https://doi.org/10.1037/cns0000137

Powell, M. B., Fisher, R. P., & Wright, R. (2005). Investigative interviewing. In N. Brewer & K. D. Williams (Eds.), *Psychology and law: An empirical perspective* (pp. 11–42). The Guilford Press.

Powell, M. B., & Snow, P. C. (2007). Guide to questioning children during the free narrative phase of an investigative interview. *Australian Psychologist*, *42*(1), 57–65. https://doi.org/10.1080/00050060600976032

Robinson, M. D., Johnson, J. T., & Robertson, D. A. (2000). Process versus content in eyewitness metamemory monitoring. *Journal of Experimental Psychology. Applied*, *6*(3), 207–221. https://doi.org/10.1037//1076-898x.6.3.207

Saraiva, R. B., Hope, L., Horselenberg, R., Ost, J., Sauer, J. D., & van Koppen, P. J. (2020). Using metamemory measures and memory tests to estimate eyewitness free recall performance. *Memory*, *28*(1), 94–106. https://doi.org/10.1080/09658211.2019.1688835

Sauer, J., Brewer, N., Zweck, T., & Weber, N. (2010). The effect of retention interval on the confidence-accuracy relationship for eyewitness identification. *Law and Human Behavior*, *34*(4), 337–347. https://doi.org/10.1007/s10979-009-9192-x

Sauer, J. 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*(3), 147–165. https://doi.org/10.1037/law0000203

Schneider, W., Visé, M., Lockl, K., & Nelson, T. (2000). Developmental trends in children's memory monitoring - evidence from a judgment-of-learning task. *Cognitive Development*, *15*, 115–134. https://doi.org/10.1016/S0885-2014(00)00024-1

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

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

Spearing, E. R., & Wade, K. A. (2022). Long retention intervals impair the confidence–accuracy relationship for eyewitness recall. *Journal of Applied Research in Memory and Cognition*, *11*(3), 384–391. https://doi.org/10.1037/mac0000014

Tenney, E. R., MacCoun, R. J., Spellman, B. A., & Hastie, R. (2007). Calibration trumps confidence as a basis for witness credibility. *Psychological Science*, *18*(1), 46–50. https://doi.org/10.1111/j.1467-9280.2007.01847.x

Tomes, J. L., & Katz, A. N. (2000). Confidence–accuracy relations for real and suggested events. *Memory*, *8*(5), 273–283. https://doi.org/10.1080/09658210050117708

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

Wells, G. L., & Bradfield, A. L. (1998). "Good, you identified the suspect": Feedback to eyewitnesses distorts their reports of the witnessing experience. *Journal of Applied Psychology*, *83*(3), 360–376. https://doi.org/10.1037/0021-9010.83.3.360

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

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*(1), 10–65. https://doi.org/10.1177/1529100616686966

How to cite this article: Caso, A., Gabbert, F., & Dando, C. J. (2024). Eyewitness confidence in the interviewing context: Understanding the impact of question type and order. *Applied Cognitive Psychology*, *38*(3), e4197. https://doi.org/10.1002/acp.4197