Eyewitness confidence : the relation between accuracy and confidence in episodic memory

Odinot, G.

3

Repeated suggestive questioning, accuracy, confidence and consistency in eyewitness event memory*

* This chapter is submitted as: Odinot, G., Wolters, G., & Giezen van, A. E. (submitted). Repeated suggestive questioning, accuracy, confidence and consistency in eyewitness event memory.

Summary

In legal practice, both confidence and consistency of the testimony of eyewitnesses are used as indicators for accuracy. The present study was designed to assess the effects of repeatedly asking correct and suggestive questions on accuracy, confidence and consistency in recall of an episodic memory. Witnesses viewed a video and answered a questionnaire containing questions that were correct and suggestive in nature. The initial cued recall test was given after 1 week and was repeated after 3 and 5 weeks. Of the questions containing suggestive information, almost half were answered with incorrect details. For the questions correct in nature, correlations between accuracy, confidence and consistency were determined for final recall. Confidence appeared to be a better predictor for accuracy than consistency. Inconsistencies consisted mainly of omitted or committed units of information that were almost as accurate as consistently recalled units. We did not find any evidence for confidence inflation with repeated questioning.

Introduction

A substantial body of research on memory for everyday events has made it abundantly clear that memory is not only fallible, i.e., prone to forgetting, but also malleable, i.e., prone to errors (e.g., Deffenbacher, 1991; Wells & Loftus, 2003). Misleading post-event information or suggestive questioning can lead to inaccuracies in eyewitness memory (e.g., Loftus, 1980, 2005). People can be led to develop even completely false memories, i.e., brought to believe they remember events that never happened (e.g., Ceci, Huffman, Smith, & Loftus, 1994; Hyman, Husband, & Billings, 1995).

To distinguish between correct and incorrect memories people often rely on expressions of confidence, or on the consistency of multiple reports of the same memory. In the study reported here, we will examine the relationship between confidence, consistency and accuracy of repeatedly probed memories of a complex event.

It is generally believed that confidence about a memory is strongly related to the accuracy of a memory. This belief is not only held by laypeople but also by members of the legal profession (Cutler, Penrod, & Stuve, 1988; Leippe, 1980; Lindsay, Wells, & O'Connor, 1989; Luus & Wells, 1994; Penrod & Cutler, 1995). The confidence expressed by an eyewitness regarding his or her testimony is a strong determinant of the perceived credibility of the eyewitness (Leippe, Manion, & Romanczyk, 1992; Lindsay et al., 1989). Studies on the relationship between accuracy and confidence, however, have found low correlations in person identification tasks (e.g., Deffenbacher, 1991; Penrod & Cutler, 1995), and modest correlations in event recall (Robinson & Johnson, 1996). Among experts, therefore, it is generally concluded that confidence is not a reliable predictor of memory accuracy (e.g., Leippe & Eisenstadt, 2007; Shaw, McClure, & Dykstra, 2007).

The relationship between accuracy and confidence may further deteriorate when misleading information is presented, either by giving misinformation after witnessing an event, or by suggestive questioning. Shaw, Garcia, & Robles (1997) showed that misinformation indeed can interfere with the accuracy-confidence relation. In three studies, participants first viewed a videotape of a simulated robbery. Afterwards, they read a post-event narrative that contained consistent, inconsistent or neutral information with respect to details in the video. In all studies participants showed overconfidence during the final memory test in all conditions. The overconfidence was, however, significantly higher in the consistent and inconsistent conditions than in the neutral condition. Similar results were reported by Ryan and Geiselman (1991). Their findings indicated that the participants were more confident about biased knowledge (i.e., answers to both leading and misleading post-event information), than about their memories of unbiased information. The effects of misleading information can be enhanced by multiple exposures. Zaragoza and Mitchell (1996) showed that participants in a three-exposure condition were significantly more likely to falsely remember the misinformation with high confidence than participants in a single-exposure condition.

The effect of suggestive questioning on accuracy and confidence of memory was studied by Shaw, Garven and Wood (1997). They reported that although answering misleading questions sometimes resulted in incorrect answers, these answers were given with less confidence. Also Roebers (2002) has reported that confidence in the responses on misleading questions was lower than in the responses for correct questions. Gerrie, Belcher and Garry (2006) asked participants about missing action details in a video. In a recognition memory test, participants confidently, but falsely, remembered some of the suggested but missing action details. In a second experiment the missing details were either crucial or not crucial for the event. In this situation participants were more likely to falsely recognize, and to be more confident, with missing noncrucial than missing crucial information. Repeating suggestive questions seems to augment its effect on confidence. Blagrove and Akehurst (2000) questioned their participants twice. The repetition led to significant increases of confidence in suggested responses to the misleading questions. Recently, also Pezdek, Sperry and Shana (2007) reported that answering the same suggestive questions twice led to an increase in confidence.

Distortions of accuracy, confidence and consistency may also occur simply by repeated questioning (or repeated reflective thought). For instance, several authors have suggested that repeated recall may cause confidence inflation because it enhances ease of retrieval or response fluency (e.g., Robinson, Johnson, & Robertson, 2000; Shaw, McClure, & Wilkens, 2001). In a recent review Shaw et al. (2007) concluded that repeated questioning generally leads to increases in the confidence ratings. These conclusions are, however, mainly based on testing memory with recognition tasks. In contrast, two other studies using cued recall for testing memory of details of complex naturalistic events failed to find evidence of confidence inflation with repeated recall of the same information (Ebbesen & Rienick, 1998; Odinot & Wolters, 2006; Turtle & Yuille, 1994).

In a crime investigation, witnesses are often questioned more than once. One of the reasons to question witnesses several times is the idea that witnesses may provide additional information during follow-up questioning. Roediger, McDermott and Goff (1997) concluded, however, that repeated recall can have both facilitating and detrimental effects on later retention. To understand these effects of repeated recall it is important to note that retrieval is not a neutral process, which leaves memory unaffected. Rather, probing memory and (re) activating memory traces is itself a learning experience. It is an active process that selectively strengthens or alters the contents of memory thus irrevocably affecting future retention (Bjork, 1975).

One characteristic of repeated recall is that it enhances the chances that various kinds of inconsistencies occur. Information remembered the first time may not be remembered later (omission error), or vice versa (commission error). Or information remembered the first time may be remembered differently a second time (distortion error). Especially the latter, but to some extend also the former types of inconsistencies in testimonies are considered as strong indicators for inaccuracies (Brewer, Potter, Fisher, Bond, & Luszcz, 1999; McNally, 2003; Talarico & Rubin, 2003). When eyewitnesses provide inconsistent information, concerns arise about the overall accuracy of the witnesses account. Fisher and Cutler (1995) have reported data from a survey, confirming that judges and lawyers strongly believe that inconsistency is predictive for inaccuracy. Conversely, consistency is often taken interpreted as an indicator or a proxy for accuracy. Accuracy and consistency, however, refer to different concept; reports of a witness can be consistent, without necessarily being accurate. Studies on the consistency-accuracy relationship are sparse. Brewer et al. (1999), Fisher et al. (1995) and Smeets, Candel and Merckelbach (2004) concluded that consistency is not a strong predictor of the accuracy of testimonies. Contrasting results are reported by Van Giezen, Arendsman and Spinhoven (2007), who found that consistency of reports for both neutral and emotional stimuli were significantly correlated with memory accuracy.

The present study was designed to assess the effects of repeatedly asking correct and suggestive questions on confidence, accuracy and consistency in the recall of an episodic memory. To that end, participants were shown a videotape of an extended complex event. Subsequently they were asked to recall as much as possible in a cued recall task and to rate their confidence in the accuracy of the answer. The recall cues consisted of open-ended questions that needed not to be answered if the participant did not remember. The suggestive questions asked about information that might have been present but was not actually shown in the video. The initial test was given after 1 week and was repeated after 3 and 5 weeks. To get more insight in delay and repetition effects, control groups received the test twice (after 3 and 5 weeks) or only once (after 5 weeks). Relations between accuracy, confidence and consistency were determined for final recall in condition 1 (three repetitions) and 2 (two repetitions).

Method

Participants

A group of 62 undergraduate students (52 female and 10 male) were recruited through publication board announcements and by a computerized sign-up system. Participants were randomly assigned to one of three conditions. All received either course credits or were paid between 10 and 20 Euros, depending on the experimental condition in which they participated.

Design

The participants were randomly assigned to the three conditions. Condition 1 (N = 21) consisted of three recall sessions 1, 3 and 5 weeks after the video presentation, Condition 2 (N = 20) consisted of two recall sessions 3 and 5 weeks after the video presentation. Condition 3 (N = 21) had only one recall session, 5 weeks after the video presentation.

Materials

Videotape. A 21 minutes long videotape, previously broadcasted on the Dutch television, was shown individually to the participants on a high quality 17-inch computer screen. The video depicts two storylines; one of a man who is helping a neighbor to get some things from a shop, and the other of a young man who recently received a motor-bike for his birthday. The two storylines converge in an accident between the car and the motorbike at the end of the video. None of the participants indicated that they had seen the video before.

Questionnaires. For the recall sessions, a questionnaire was constructed consisting of 28 open-ended questions of which five contained suggestive information (see appendix 1). The questionnaire started with a very general question asking the participants to describe the two story lines in general terms. This question was included to reinstate and refresh the memory of the video before proceeding with the more specific questions. All other questions were open-ended question asking about specific details shown in the video. Some questions were more specific (e.g., "describe the car of the man") than others (e.g., "give a full description of the accident and try to be as complete as possible"). Five questions contained suggestive information, asking about aspects that were not shown in the video. For instance: "The driver of the car that was involved in the accident was injured. Where and how bad was he injured?", while the driver of the car was not injured at all. These five questions were randomly interspersed with the questions that were correct in nature. The same questionnaire was used in all recall sessions.

Procedure

During the first session, participants watched the video individually on a computer monitor. They were told to pay attention because they would have to recall the event later. In the

initial and subsequent recall sessions participants were instructed to try to recall information from the original video. They were told to imagine that they were the only witnesses, and that it was important therefore to report as accurate as possible about any details they remembered from the original video-presentation. It was also stressed, however, that if they could not remember the answer from the video, they should refrain from answering by indicating "do not know".

To allow a fine-grained analysis of the recall data, participants were instructed to write the answers to the questions in small units of information. A unit was described as a single element or aspect of information. To explain this to the participants, the following example was given; *question;* 'What did the dog do when it came out of the water?' *answer;* 'it climbed on the bank'; 'it shook off the water'; 'it ran to his boss'. To encourage the subjects to give single elements of information, the lines on the answering sheet were restricted in length. Participants could answer with as many units of information as they needed. After writing down the answer, participants were asked to indicate their confidence regarding the accuracy of each unit of information given on a 7-point scale (1= very uncertain, 7 = absolutely certain).

After finishing the questionnaire, the experimenter judged if the information was given in small units and if confidence indications were made to every unit of information. Only in a few cases, participants were asked to do so afterwards. After completing the final session, participants were debriefed and paid or given credits.

The correct and suggestive questions were scored separately. All units of information provided by the participants on the correct questions were scored as correct or incorrect. Information was scored *correct* when it corresponded with information from the video. *Incorrect* information consists of units of information not present in the video, which were apparently incorrectly remembered or fantasized by the participant. The answers given on questions containing suggestion were scored in three categories. First, participants could avoid an answer by choosing the 'do not know' option. Second, the participant could respond with an answer in which the suggestion was rejected explicitly (e.g., "I haven not seen any blood"), this was scored as *correct*. Third, any other answer indicating acceptance of the suggestion (e.g., "there was blood on his arm") was scored as *incorrect*.

The responses were also scored on consistency by comparing the units of information for each participant across the recall sessions. Information was scored as *consistent* when it was recalled in all subsequent recall sessions (three times in condition 1 and two times in condition 2). Information did not have to be literally the same to be consistent, for instance, information as "the car didn't stop", provided during the first session and subsequently recalled as "the car kept moving" during the next recall session was scored as consistent. Information not recalled across all recall sessions or recalled differently was scored as *inconsistent*.

All answers were scored on accuracy and consistency independently by two experimenters. Their scoring coincided in 94% of the cases. In case of disagreement, a third experimenter settled the dispute. Of all units generated 1.8% could not be classified as correct or incorrect; these units of information were discarded from further analysis. The mean number of information-units given to correct questions was 2.53 (with a minimum of 1 and a maximum of 29).

Results

There are five main foci in this study: responses to suggestive questions, responses to correct questions, the consistency and confidence in the responses, and the relationship between accuracy, confidence and consistency. We will discuss these foci in separate sections.

Suggestive Questions

The questionnaire contained five questions that were suggestive in nature. Basically there were three ways to respond: withholding a response by using the "I don't know" option, giving a correct response in which the suggestion is rejected ("I haven't seen any blood"), or giving an incorrect response in which the suggestion is accepted and results in a fantasized answer ("The man was bleeding at his arm"). Results were analyzed by comparing the numbers (or proportions) of each type of answer as a function of delay and repetition (see Table 1).

Retention interval. To analyze the effects of retention interval, the first recall sessions in the 1-, 3-, and 5-week interval conditions were compared and tested for differences in accuracy and confidence. Note that this is a comparison between groups.

The proportion of suggestive questions answered with an "I do not know" response (0.46 after 1 week, 0.35 after 3 weeks and 0.52 after 5 weeks), did not differ significantly as a function of delay (F (2, 59) = 2.21, *NS*). However, the number of responses in which the suggestion was correctly rejected by the participants, showed a significant decrease with longer retention intervals (F (2, 59) = 5.34, $p < .05$). The proportion of these answers dropped from 0.23 after a retention interval of one week, to 0.11 and 0.10, after three and five weeks, respectively. The confidence in these correct responses did not show a significant decrease with retention interval. The mean confidence was 5.64, 5.30 and 5.28, after 1, 3 and 5 weeks, respectively (F (2, 34) = 0.30, *NS*).

Table 1 Proportions of correctly rejected (C), incorrectly fantasized (I) and "Do not know" (D) answers provided on the suggestive questions, and corresponding average confidence ratings (sd in parentheses), as a function of retention interval and repeated recall.

Suggestive questions		Retention interval								
		1 week			3 weeks			5 weeks		
		C	I	D	C	I	D	C	I	D
Condition 1	Proportion	.23	.30	.46	.20	.34	.46	.20	.34	.46
(N=21)	Confidence	5.64 (1.2)	4.23 (1.4)		5.84 (.85)	3.92 (1.4)		5.90 (1.3)	4.35 (1.4)	
Condition 2	Proportion				.11	.54	.35	.10	.59	.31
(N=20)	Confidence				5.30 (1.7)	3.85 (1.4)		4.88 (1.88)	3.82 (1.3)	
Condition 3	Proportion							.09	.38	.52
(N=21)	Confidence							5.28 (1.1)	3.89 (1.6)	

The proportion of questions that were answered incorrectly was 0.30 after 1 week, 0.54 after 3 weeks and 0.38 after 5 weeks. These differences were significant (F (2, 59) = 4.04, p < .05), but there is no clear pattern relating incorrect answers and delay. Bonferroni post-hoc test showed a significant difference only between the interval of one and three weeks (p < .05). The mean confidence in these incorrect answers (4.23, 3.85 and 3.89, after 1, 3 and 5 weeks, respectively) showed a decreasing, but not significant, trend (F (2, 50) = .35, NS).

Repeated recall. The effect of repeated recall on accuracy and confidence is analyzed as a within subjects repeated measures variable, for condition 1 (comparing initial recall after 1 week with repeated recall after 3 and 5 weeks) and 2 (comparing initial recall after 3 weeks with repeated recall after 5 weeks). Repetition of recall did not affect the proportion of "do not know" responses. In condition 1 the proportions were 0.46, 0.46 and 0.46 for recall sessions 1, 2 and 3, respectively (F (2, 19) = .00, *NS*), and in condition 2 these proportions were 0.35 and 0.31, in recall sessions 1 and 2, respectively (t (19) = 1.01, *NS*). Also the proportions of correct rejection responses, and the mean confidence in these responses, were not significantly affected by repetition.

The effect of repetition on the proportion of incorrect answers showed a small, and insignificant increase, both in condition 1 (0.30, 0.34 and 0.34) and condition 2 (0.54 and 0.59). The mean confidence in the incorrect responses in condition 1 (4.15, 4.08 and 4.76 in the first, second and third recall session, respectively) showed a significant, but somewhat difficult to explain, difference between the second and third recall (F (2, 12) = 6.15, p < .05, Bonferroni p < .05). The mean confidence of incorrect answers in condition 2 did not show an effect of repetition (M = 3.85 and 3.82, respectively, t (16) = .14, *NS*).

It is of particular interest to compare the proportions and average confidence of correct (suggestion rejected) and incorrect (suggestion accepted) responses. From Table 1 it can be inferred that the proportion of incorrect answers was larger than the proportion of correct answers in all conditions. It is also clear, however, that the mean confidence of incor-

rect answers was lower than the confidence of correctly answered questions. Bonferroni posthoc tests showed that all these differences in proportions and average confidence were significant (p < .05). Although on average confidence in incorrect answers is lower than in correct answers, still quite a few incorrect responses were given with a high level of confidence. The distribution of correct and incorrect answers to suggestive questions over confidence ratings is shown in Table 2.

Overall the results on the questions containing suggestive information indicate that suggestion had a substantial effect. In the first recall session, the subjects in conditions 1, 2 and 3 answered (either with an "I do not know" or a rejection answer) in 69%, 46% and 62% of the cases without falling for the suggestion, respectively. This means, however, that in 31%, 54% and 38% of the cases an incorrect answer was given, even when it was explicitly allowed to refrain from answering. Moreover, although on average confidence in incorrect answers was lower than for correct answers, a substantial proportion was given with a high level of confidence.

Table 2 The number of correct and incorrect responses to suggestive questions at each confidence level as a function of conditions; the proportions of incorrect answers are presented in parentheses.

Condition	Retention interval	Confidence Scale							Number of Do not know responses
		1	2	3	4	5	6	7	
Condition 1 N= 21	1 week	3 (.66)	3 (1.0)	9 (.88)	6 (.50)	13 (.46)	12 (.58)	10 (.20)	49
	3 weeks	5 (1.0)	4 (1.0)	5 (1.0)	7 (.86)	14 (.57)	11 (.45)	10 (.20)	48
	5 weeks	3 (1.0)	5 (1.00	4 (.75)	8 (.75)	12 (.83)	10 (.60)	15 (.26)	48
Condition 2 N=20	3 weeks	7 (.78)	9 (.88)	11 (.90)	11 (.91)	9 (.77)	10 (.70)	8 (.63)	35
	5 weeks	4 (1.0)	10 (.80)	11 (1.0)	13 (1.0)	16 (.81)	11 (.73)	5 (.60)	31
Condition 3 N= 21	5 weeks	5 (1.0)	7 (1.0)	4 (1.0)	11 (.73)	9 (.55)	9 (.88)	5 (.60)	55

The 20 participants in condition 2 seem to be more prone to suggestion than the participants in conditions 1 and 3. Because there are no obvious differences in the presentation and testing conditions to explain this difference, it suggests that it may be caused by group differences. Therefore, we also looked at individual results. The individual participants, indeed, showed a wide variation in their responses to the suggestive questions. Of the 62 participants, nine refrained from answering any of the suggestive questions during first recall, whereas 10 participants answered at least four of the five suggestive questions. As the results already suggested, relatively more participants in condition 2 than in the other two conditions answered incorrectly to suggestive questions.

Correct Questions

To analyze the effects of retention interval and repeated recall on accuracy and confidence, we determined the number of correct and incorrect units of information given by the participants in all conditions. The proportions correct and incorrect units of information and the corresponding mean confidence judgments are shown in Table 3.

First, we examined the effect of retention interval on the number of questions that were answered with an "I don't know" response and on the total number of units of information. During the first recall session after one week only 5.4% of the answers to all 23 correct questions were "I don't know" responses. After retention intervals of 3 and 5 weeks, this number was not significantly different (4.4 % and 7.9 %, respectively; (F (2, 59) = 1.40, *NS*). The average total number of units of information provided per participant to the 23 correct questions also did not show a significant effect of delay (F (2, 59) = 1.16, *NS*). The average total number of units of information, correct or incorrect, that was recalled after 1, 3 and 5-week intervals were 61.7, 56.2 and 55.9, respectively.

Table 3 Proportions correct and incorrect units of information, and corresponding average confidence ratings (sd in parentheses), as a function of retention interval and repeated recall.

Correct questions		**Retention interval**					
		1 week		**3 weeks**		**5 weeks**	
		Correct	**Incorrect**	**Correct**	**Incorrect**	**Correct**	**Incorrect**
Condition 1	Proportion	.88	.12	.85	.15	.85	.15
N=21	Confidence	6.37 (.66)	5.50 (.68)	6.34 (.53)	5.75 (.72)	6.46 (.44)	5.55 (.84)
Condition 2	Proportion			.80	.20	.79	.21
N=20	Confidence			6.25 (.43)	5.12 (.93)	6.20 (.53)	5.23 (1.01)
Condition 3	Proportion					.77	.23
N=21	Confidence					6.10 (.47)	4.95 (.48)

Retention interval. The effect of retention interval was analyzed by comparing results of the first recall sessions only. The number of correctly recalled units of information showed a significant decrease with retention interval, F (2, 59) = 21.99, $p < 0.001$. Bonferroni post-hoc tests showed significant differences between 1- and 3-week intervals (proportions correct 0.88 and 0.80, respectively, $p < .05$), and between 1- and 5- week intervals (proportions correct 0.88 and 0.77, respectively, $p < .05$). Confidence in the correct units of information also seems to decrease with longer intervals, but the difference after one, three and five weeks (M = 6.37, 6.25 and 6.10, respectively) was not significant (F (2, 59) = 1.33, *NS*).

The number of incorrectly recalled units of information showed a significant increase with retention interval (F (2, 59) = 21.99, $p < .001$). Proportion incorrect after one week (0.12) was significantly lower than after three (0.20) or five weeks (0.23). Confidence in the incorrect units showed a significant decrease with longer intervals (F (2, 59) = 3.22, $p < .05$). Bonfer-

roni post-hoc tests indicated that the mean level of confidence for incorrect recalled information was higher after 1 week (M = 5.50) than after 5 weeks (M = 4.95, $p < .05$).

In sum, the proportion of correctly recalled units of information decreased with longer retention intervals, while at the same time the proportion of incorrect information significantly increased. In the recall of both correct and incorrect units there is a trend that confidence decreases with increasing retention intervals. Interestingly, during initial recall participants were always significantly more confident about correctly than about incorrectly recalled units of information ($p < 0.01$, for all delay conditions).

Repeated recall. The effect of repeated recall was analyzed as a within subjects repeated measures variable, for condition 1 (3 recall sessions) and condition 2 (2 recall sessions). The proportion correctly recalled units of information remained almost the same across the subsequent recall sessions, both in condition 1 (0.88, 0.85 and 0.85, respectively) and in condition 2 (0.80 and 0.79, respectively). Also, the mean levels of confidence for these correctly recalled units of information were not significantly influenced by repeated recall in condition 1 (6.37, 6.34 and 6.46, respectively; F (2, 40) = .43, *NS*), nor in condition 2 (6.25 and 6.20, respectively, t (19) = .86, *NS*).

Also the proportion of incorrectly recalled units showed no effect of repeated recall (F (2, 40) = 2.25, *NS*). Proportions incorrect units in condition 1 were 0.12, 0.15 and 0.15, in the first, second and third recall session, respectively. In condition 2 these proportions were 0.20 and 0.21, respectively. Also the mean confidence of the incorrectly recalled units was unaffected by repetition, both in condition 1 (5.50, 5.75 and 5.55 respectively; F (2, 40) = 1.29, *NS*), and in condition 2 (5.12 and 5.23, respectively, t (19) = -.60, *NS*).

Total units of information and proportions of inaccurate units as a function of confidence judgment levels are shown in Table 4. From this Table it is clear that in all conditions the proportion of incorrect units drops steadily with higher confidence levels. After a recall interval of one week, the proportion of errors with high confidence ratings is 0.06. With intervals of 3 and 5 weeks this proportion increases to 0.11. These figures suggest that confidence may be used as a modest indicator of accuracy. However, one should always be aware that high confidence is never a guarantee for accuracy.

Table 4 The total number of units of information at each confidence level per condition for the correct questions; proportions of incorrect units are presented in parentheses.

	Retention interval	Confidence Scale							Total units of information
		1	2	3	4	5	6	7	
Condition 1 N = 21	1 week	1 (1.00)	12 (.75)	23 (.430	60 (.33)	151 (.23)	209 (.18)	813 (.06)	1269 (.12)
	3 weeks	6 (.33)	11 (.54)	24 (.58)	43 (.26)	136 (.25)	190 (.22)	788 (.09)	1198 (.15)
	5 weeks	3 (.67)	17 (.53)	30 (.33)	43 (.37)	119 (.29)	176 (.21)	805 (.09)	1193 (.15)
Condition 2 N = 20	3 weeks	7 (.86)	40 (.52)	32 (.56)	74 (.42)	141 (.30)	227 (.21)	603 (.10)	1124 (.20)
	5 weeks	10 (.60)	33 (.58)	44 (.45)	52 (.38)	151 (.32)	180 (.24)	540 (.11)	1010 (.21)
Condition 3 N = 21	5 weeks	16 (.31)	38 (.53)	61 (.52)	102 (.44)	148 (.34)	213 (.26)	596 (.11)	1174 (.23)

Consistency

The consistency in the responses across the recall sessions was determined for conditions 1 and 2, for the correct and suggestive questions separately.

Suggestive questions. A response to the same question was scored as consistent when the suggestion was correctly rejected across all recall sessions or when the same correct or incorrect answer was given. An answer was scored as inconsistent when an answer was not recalled consistently across the recall sessions or recalled differently.

In condition 1 a proportion of 0.74 of all answers to the suggestive questions was consistent across the three recall sessions. Of this proportion, 0.35 was a consistent "I don't know" answer, 0.18 were answers in which participants consistently rejected the suggestion, and 0.21 were recalled consistently but incorrectly. Of all answers to the suggestive questions a proportion of .26 was inconsistent. Of this proportion, 0.06 were inconsistencies in which a previously correct rejection or don't know answer changed into an incorrect answer. Mean confidence in consistent responses was significantly higher (M = 5.25) than confidence in inconsistent responses (M = 4.21; t (13) = -2.77, p < .05). Confidence did not increase significantly over repetitions (M = 5.78, 5.94 and 6.00 for recall 1, 2 and 3, respectively).

In condition 2 a proportion of 0.84 of all responses to the suggestive questions were consistent. Of this proportion, 0.26 were consistent "I don't know" responses, 0.08 were consistent correct rejection responses, and a remarkable 0.50 were consistently incorrect. Confidence was higher for all consistent responses (M = 4.63) than for the inconsistent responses (M = 3.82), but this difference just failed to reach significance (t (6) = -2.3, *NS*). Again, repetition of consistent answers did not affect confidence significantly.

Correct Questions. Consistencies in responses to the correct questions were analyzed by comparing the units of information across the recall sessions. In the responses to the correct questions in condition 1, the participants provided a total of 1683 single units of information. Almost half of all units, a proportion of 0.49, were consistently recalled during the three recall sessions. Of this proportion 0.43 was accurate and 0.06 was consistent but inaccurate. A proportion of 0.51 was answered inconsistently. In large part (0.43) these were correct units of information missing in one or two of the recalls sessions (omission or commission). Of the part that was incorrect (0.08), most were incorrect details given on one occasion but not on another (0.06), and only 0.02 were distortions (different answers on different occasions).

The proportion correct of consistently and inconsistently recalled information was the same (0.43). Also the proportion incorrect of consistently (0.06) and inconsistently (0.08) recalled information did not differ significantly (t (20) = -1.24, *NS*). So consistently recalled information is not more correct than inconsistently recalled information. The participants were however, significantly more confident about consistently recalled information (M = 6.12) than about inconsistently recalled information (M = 5.89, t (20) = 2.13, $p < .05$).

In condition 2, a total of 1416 units of information were provided to the correct questions. A portion of .51 of all these units was consistently recalled across the two recall sessions, of which 0.43 was correct and 0.08 was incorrect. Of the 0.49 inconsistently recalled units, 0.37 were correct (omissions and commissions), 0.09 changed from correct to incorrect, and 0.03 were distortions. Also here, the proportions correct and incorrect of consistently and inconsistently recalled information did not differ significantly. Although participants expressed more confidence in consistent information (M = 5.68) than inconsistent information (M = 5.51), this difference was not significant (t (19) = 1.07, NS).

In sum, in both conditions roughly half of all the information provided was consistent across the recall sessions. Interestingly, the proportion correct of inconsistently recalled units of information was about the same as in consistent recall, but participants were more confident about consistently than inconsistently recalled information.

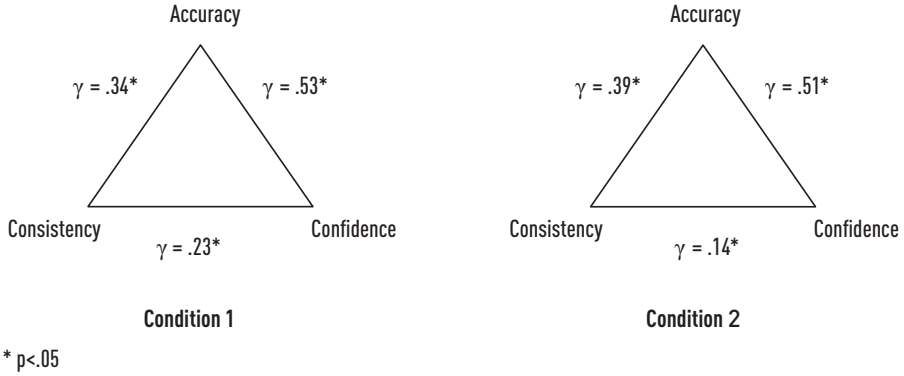
Relation between accuracy, confidence and consistency

To quantify the relations between accuracy, confidence, and consistency for the correct questions, we calculated Goodman-Kruskal gamma correlation coefficients in the responses during final recall only. All units provided during final recall in conditions 1 and 2 were scored for accuracy, confidence and consistency. A unit was scored as consistent when it was recalled across all sessions; information was scored as inconsistent when it was recalled during final recall, but not at all or differently during previous sessions. Next, accuracy and confidence was determined for all consistent and inconsistent units to calculate gamma correlation coefficients, shown in Figure 1.

As can be inferred from this figure all correlations were positive and significant ($p < .05$). However, the relation between confidence and accuracy is stronger than consistency and accuracy. Relatively high correlations between accuracy and confidence were also found in the first and second recall in condition 1 ($\gamma = .61$ and .51, $p < .05$), the first recall in condition 2 ($\gamma = .55$, $p < .05$), and in condition 3 ($\gamma = .54$, $p < .05$).

The rather modest correlations between consistency and accuracy are in line with the previously discussed finding that accuracy rate was about the same for consistent and inconsistent answers. We also noted already that in the conditions as studied here, inconsistencies are mainly errors of omission or commission, i.e., incomplete recall in subsequent sessions. The results clearly indicate that this kind of incompleteness is not related to accuracy. In contrast, we found very few instances of the kind of inconsistency that would certainly predict inaccuracy, namely changes in the information provided to the same questions (distortions).

Figure 1 Gamma correlations between accuracy, confidence and consistency for the correct questions during final recall, condition 1 and 2, respectively.



* p<.05

Discussion

The purpose of this study was to examine the effect of repeated questioning and retention interval on the accuracy, confidence and consistency in the recall of a complex episodic event when memory is probed with questions that are suggestive or correct in nature.

The inclusion of questions containing suggestive information, inconspicuous interspersed with correct questions, led to particularly revealing findings. In total, almost half of the suggestive questions were answered with obviously incorrect details (depending on groups and conditions, proportions incorrect answers ranged from 0.30 and 0.59) and longer retention intervals resulted in less correct rejections of the suggestion. Moreover, a substantial proportion of the incorrect responses were consistently reported across the recall sessions.

Although on average incorrect responses were given with lower confidence, still a substantial proportion was given with moderate to high degrees of confidence.

The main findings for the correct questions were that after longer intervals before first questioning less correct units of information are recalled and with lower confidence levels. Although the number of incorrect units of information increases with longer retention intervals, incorrect answers remain a minority compared to correct answers, and they are generally characterized by lower confidence levels. Repeated questioning did not enhance recall, nor did it influence the proportion of inaccurate recall. Moreover, we did not find any evidence for confidence inflation with repeatedly retrieving identical information from memory, replicating previous findings of Odinot and Wolters (2006).

Results with regard to consistency were also revealing. Of the units of information given to correct questions only about half was mentioned consistently over subsequent recall sessions. Interestingly, the percentage correct was about the same for consistently and inconsistently recalled units of information (e.g., in condition 1 consistent answers were 88% correct and inconsistent answers were 84% correct). This finding is due to the fact that most inconsistencies were errors of omission or commission, i.e., mentioning particular units of information in one recall session but not in another. Of the 16% inconsistent and incorrect units of information, only 5% were actual distortions (changes in the content of the information provided), the rest were incorrect units of information that were given on one occasion but not on another.

The conclusion for the relationship between accuracy, confidence and consistency is that in the conditions as studied all these relations are modest. Neither confidence nor consistency is very useful for predicting accuracy, but if anything, confidence is a better indicator than consistency.

The outcome of this study has some important practical implications. First, our findings illustrate that questions asked have to be free of suggestion. Obviously, this is not a new finding, but it emphasizes once again the importance of correctly interviewing witnesses. Any suggestion of misleading information strongly increases the chance that an incorrect answer is provided, and neither confidence nor consistency can be used to detect these inaccuracies.

Second, with correct questions our results show relatively high confidence-accuracy correlations. Especially with the shortest recall interval (one week), the large majority of answers given with a high confidence tend to be correct. With some caution, therefore, in event recall confidence may be used as a partial indicator of accuracy during the early stages of an investigation.

Third, in this study consistency is not strongly related with accuracy, a finding that is in line with most studies reported on this issue (Brewer et al., 1999; Fisher & Cutler, 1995; Smeets et al., 2004). In legal practice inconsistencies are seen as strongly predictive for accuracy. Although this is obviously true for contradictory inconsistencies (distortions), it is not true

for inconsistencies due to incomplete statements (errors of omission or commission). Witnesses should not be judged as unreliable when previously given information is omitted in later statements or when in later statements additional information is provided.

Fourth, repeatedly asking the same question did not seem to inflate the confidence expressed in the answers provided. Even when identical information was repeatedly and consistently retrieved from memory, confidence was not enhanced. This is not in line with the retrieval fluency hypothesis as suggested by Shaw (1996) and Shaw and McClure (1996). According to this hypothesis, the ease with which an item can be retrieved from memory may be used for confidence judgments, with greater ease of retrieval yielding higher confidence judgments. This hypothesis is also used to explain 'imagination inflation' findings, showing that repeated imagination causes an increased tendency to judge an imagined event as an event that actually happened (Thomas & Loftus, 2002). In this study however, also no evidence for imagination inflation is found with the repetition of fantasized responses.

The general belief among memory experts is that we should be reluctant to use confidence as an indication of accuracy. This might be true for line-up identifications, but probably not for the recall of events. A witness who is very confident remembering some detail of a well-observed event is much more likely to be correct than incorrect, provides an important instrument. However, even a highly confident witness is likely to err in about 10% of the details provided. It is important, therefore, to be reluctant in applying the knowledge about the relationship between accuracy and confidence. Firstly, this knowledge is better used during police investigations, when possible errors are not crucial, than using it as evidence during a court trial. Secondly, our study shows that confidence as an indicator for accuracy becomes less strong after longer retention intervals. This is again an argument to restrict the use of confidence to the early stages of a police investigation and not to use it in later stages, i.e., in front of a judge or jurors. The time between witnessing an event and telling about it in court is much longer than the 5 weeks tested in this experiment. Moreover, with an increasing period between witnessing an event and testifying about it, there is an increasing possibility that a witness is exposed to suggestion and misinformation. As our results show, this clearly results in a less reliable relation between accuracy and confidence. Therefore, the conclusion remains that judges, lawyers and prosecutors have to be aware that the confidence expressed by a witness in court about their memories of an event is not a good predictor for accuracy.

References

Bjork, R. A. (1975). Retrieval as a modifier. In R. L. Soslo (Ed.), *Information processing and cognition: The Loyola symposium* (pp. 123-144). Hillsdale, NJ: Erlbaum.

Blagrove, M., & Akehurst, L. (2000). Effects of sleep loss on confidence-accuracy relationships for reasoning and eyewitness memory. *Journal of Experimental Psychology: Applied, 6*, 59-73.

Brewer, N., Potter, R., Fisher, R. P., Bond, N., & Luszcz, M. A. (1999). Beliefs and data on the relationship between consistency and accuracy of eyewitness testimony. *Applied Cognitive Psychology, 13*, 297-313.

Ceci, S. J., Huffman, M. L. C., Smith, E., & Loftus, E. F. (1994). Repeatedly thinking about a non-event: Source misattribution among preschoolers. *Consciousness and Cognition, 3*, 388-407.

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. (1991). A maturing of research on the behavior of eyewitnesses. *Applied Cognitive Psychology, 5*, 377-402.

Ebbesen, E. B., & Rienick, C. B. (1998). Retention interval and eyewitness memory for events and personal identifying attributes. *Journal of Applied Psychology, 83*, 745-762.

Fisher, R. P., & Cutler, B. L. (1995). The relation between consistency and accuracy of eyewitness testimony. In G. Davies & S. lloyd Bostock & M. McMurran & C. Wilson (Eds.), *Psychology, law and criminal justice: International developments in research and practice* (pp. 21-28). Berlin: De Gruyter.

Gerrie, M. P., Belcher, L. E., & Garry, M. (2006). 'Mind the gap': False memories for missing aspects of an event. *Applied Cognitive Psychology, 20*, 689-696.

Giezen van, A. E., Arensman, E., & Spinhoven, P. (2007). Memory for emotionally arousing and neutral stimuli: An experimental study on accuracy and consistency of memory in victims of violent crime. *Manuscript submitted for publication*.

Hyman, I. E., Husband, T. H., & Billings, J. F. (1995). False memories of Childhood experiences. *Applied Cognitive Psychology, 9*, 181-197.

Leippe, M. R. (1980). Effects of integrative memorial and cognitive processes on the correspondence of eyewitness accuracy and confidence. *Law and Human Behavior, 4*, 261-274.

Leippe, M. R., & Eisenstadt, D. (2007). Eyewitness confidence and the confidence-accuracy relationship in memory for people. In R. C. L. Lindsay & D. F. Ross & J. D. Read & M. P. Toglia (Eds.), *Handbook of eyewitness psychology: Memory for people* (Vol. 2, pp. 377-425). Mahwah, New Jersey: Lawrence Erlbaum Associates, publishers.

Leippe, M. R., Manion, A. P., & Romanczyk, A. (1992). Eyewitness persuasion: How and how well do fact finders judge the accuracy of adults' and children's memory reports? *Journal of Personality and Social Psychology, 63*, 181-197.

Lindsay, R. C. L., Wells, G. L., & O'Connor, F. J. (1989). Mock-juror belief of accurate and inaccurate eyewitnesses. *Law and Human Behavior, 13*, 333-339.

Loftus, E. F. (1980). *Memory: Surprising new insights into how we remember and why we forget*. Addison-Wesley: Publishing Company, Inc.

Loftus, E. F. (2005). Planting misinformation in the human mind: A 30-year investigation of the malleability of memory. *Learning & Memory, 12*, 361-366.

Luus, E. C. A., & Wells, G. L. (1994). Eyewitness identification confidence. In D. F. Ross & J. D. Read & M. P. Toglia (Eds.), *Adult Eyewitness Testimony* (pp. 348-361). Cambridge: Cambridge University Press.

McNally. (2003). *Remembering trauma*. Cambridge, MA: The Belknap Press of Harvard University Press.

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

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

Pezdek, K., Sperry, K., & Shana, O. M. (2007). Interviewing witnesses: The effect of forced confabulation on event memory. *Law and Human Behavior, 31*, 463-478.

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

Robinson, M. D., Johnson, J. T., & Robertson, D. A. (2000). Process versus content in eyewitness metamemory monitoring. *Journal of Experimental Psychology: Applied, 6*, 207-221.

Roebers, C. M. (2002). Confidence judgments in children's and adults' events recall and suggestibility. *Developmental Psychology, 38*, 1052-1067.

Roediger, H. L., McDermott, K. B., & Goff, L. M. (1997). Recovery of true and false memories: Paradoxical effects of repeated testing. In M. A. Conway (Ed.), *Recovered memories and false memories* (pp. 118-149). Oxford: Oxford University Press.

Ryan, R. H., & Geiselman, E. R. (1991). Effects of biased information on the relationship between eyewitness confidence and accuracy. *Bulletin of the Psychonomic Society, 29*, 7-9.

Shaw, J. S. (1996). Increases in eyewitness confidence resulting from postevent questioning. *Journal of Experimental Psychology: Applied, 2*, 126-146.

Shaw, J. S., Garcia, L. A., & Robles, B. E. (1997). Cross-language postevent misinformation effects in spanish-english bilingual witnesses. *Journal of Applied Psychology, 82*, 889-899.

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

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.

Shaw, J. S., McClure, K. A., & Dykstra, J. A. (2007). Eyewitness confidence from the witnessed event through trial. In M. P. Toglia & J. D. Read & D. F. Ross & R. C. L. Lindsay (Eds.), *Handbook of eyewitness psychology* (Vol. 1, pp. 371-397). Mahwah, New Jersey: Lawrence Erlbaum Associates, Publishers.

Shaw, J. S., McClure, K. A., & Wilkens, C. E. (2001). Recognition instructions and recognition practice can alter the confidence-response time relationship. *Journal of Applied Psychology, 86*, 93-103.

Smeets, T., Candel, I., & Merckelbach, H. (2004). Accuracy, completeness, and consistency of emotional memories. *American Journal of Psychology, 117*, 595-609.

Talarico, J. M., & Rubin, D. C. (2003). Confidence, not consistency, characterizes flashbulb memories. *Psychological Science, 14*, 455-461.

Thomas, A. K., & Loftus, E. F. (2002). Creating bizarre false memories through imagination. *Memory & Cognition, 30*, 423-431.

Turtle, J. W., & Yuille, J. C. (1994). Lost but not forgotten details: Repeated eyewitness recall leads to reminiscence but not hypermnesia. *Journal of Applied Psychology, 79*, 260-271.

Wells, G. L., & Loftus, E. F. (2003). Eyewitness memory for people and events. In A. M. Goldstein (Ed.), *Handbook of psychology: Forensic psychology* (Vol. 11, pp. 149-160). New York: John Wiley and Sons, Inc.

Zaragoza, M. S., & Mitchell, K. J. (1996). Repeated exposure to suggestion and the creation of false memories. *Psychological Science, 7*, 294-300.

Appendix 1

The suggestive questions

"One of the boys tried to impress the girls by riding circles around them on his moped and making a lot of noise with the engine. What was the irritated reaction of the two girls?"

In the video was shown that the two girls had to stop their bicycles because the boy was driving in circles around them on his moped, making a lot of noise. But the girls did not say anything, giggled a little bit and followed their way. There was no irritation shown by the girls.

"At a certain moment two men had a disagreement while they were trading a car. What was this disagreement about?"

In the video was shown that two men were trading a moped. The atmosphere between the men was friendly and there was no disagreement at all.

"The minivan of the father was partly plastered with an advertisement. Can you describe the text and/ or illustration?"

The minivan of the father was not covered with advertisement. However, this kind of minivan is often used for business purposes and advertisements.

"The driver of the car that was involved in the accident was injured. Where and how bad was he injured?"

The driver of the car was not injured at all. In the video was shown that the driver stepped out of his car, right after the accident to check the condition of the person who lies in front of the car.

"The victim of the accident was seriously injured. Where was the boy bleeding?"

In the video was shown that the boy was driving his moped and hit the back of the car in full speed. The boy flies over the car and rolls back on the ground. He lies unconsciously in front of the car however, he is not bleeding or visually wounded.