Testing promotes eyewitness accuracy with a warning: Implications for retrieval enhanced suggestibility

Ayanna K. Thomas [a,*], John B. Bulevich [b], Jason C.K. Chan [c]

[a] Department of Psychology, 490 Boston Ave., Tufts University, Medford MA, United States
[b] Department of Psychology, Rhode Island College, Providence, RI, United States
[c] Department of Psychology, W112 Lagomarcino Hall, Iowa State University, Ames, IA 50011-3180, United States

ARTICLE INFO

ABSTRACT

Numerous studies have demonstrated that repeated retrieval boosts later retention. However, recent research has shown that testing can increase eyewitness susceptibility to misleading post-event information (e.g., Chan, Thomas, & Bulevich, 2009). The present study examines the effects of warning on this counterintuitive finding. In two experiments, subjects either took an initial test or performed a filler task after they viewed a video event. They were then given post-event information before they took a final test. Critically, one group of subjects was warned about potential inaccuracies in the post-event narrative and the other group was not. Without a warning, subjects who received an initial test were more likely to endorse misleading post-event information, replicating the retrieval-enhanced suggestibility (RES) effect. However, this RES effect was eliminated when subjects were warned about the veracity of the narrative. These results are consistent with a retrieval fluency account of RES.

© 2010 Elsevier Inc. All rights reserved.

Introduction

In their classic study, Loftus, Miller, and Burns (1978) demonstrated that exposure to misleading information after witnessing an event reduced accuracy on a later memory test. Variants of this general finding have since been demonstrated in dozens of papers. The relevant societal implication of eyewitness fallibility has encouraged an investigation into techniques that could be employed to resist effects of misleading post-event information. Recently, Chan et al. (2009) attempted to reduce eyewitness suggestibility by testing subjects prior to the presentation of a post-event narrative. The logic was that initial testing would reduce people's susceptibility to later misinformation because the initial test would enhance memory for the original event. This hypothesis was based on the well established testing effect (for a review, see Roediger & Karpicke, 2006), which is the finding that taking an intervening test between learning and a final delayed test boosts performance on that final test. Contrary to this hypothesis, Chan et al. found that subjects who received an initial test were less accurate on a final test of memory, and more likely to endorse misleading post-event information, than those who received only the final test. In this paper, we refer to this finding as retrieval-enhanced suggestibility (RES).[1]

In the present study, we investigated whether retrieval fluency of the misinformation accounts for the increased suggestibility that occurs under repeated retrieval conditions.

---

* Corresponding author.
  E-mail address: ayanna.thomas@tufts.edu (A.K. Thomas).

[1] In the Chan et al. (2009) paper, this finding was referred to as the "reversed testing effect." However, upon further considerations, we feel that this terminology is not representative of the most important aspect of the finding – that initial retrieval can increase eyewitness suggestibility to misinformation. Further, Chan and Langley (in press) have reported that a regular testing effect can co-occur with retrieval-enhanced suggestibility (RES), thus, we feel that RES is a more suitable and descriptive term of this finding.

In the RES procedure, after viewing a complex video event, subjects take a test and then are presented with a post-event narrative, which includes details associated with the initially tested material. We hypothesize that those details in the narrative may capture attention, and are thus better encoded (for a similar finding in verbal learning, see Robbins & Irvin, 1976; Tulving & Watkins, 1974). Further, this enhanced encoding of the misinformation increases its ease of retrieval later, which is manifested as increased susceptibility to misinformation (i.e., RES). In the present paper, we refer to this account as the retrieval fluency hypothesis (Baddeley, 1982a; Jacoby & Dallas, 1981; Jacoby, Kelley, & Dywan, 1989b). The term 'retrieval fluency' refers to the ease with which a piece of information is recalled from memory (Baddeley, 1982b; Jacoby & Dallas, 1981; Jacoby, Kelley, & Dywan, 1989a). This retrieval fluency hypothesis depends on two propositions: (a) initial testing enhances learning of the post-event information, which increases its ease of retrieval, and (b) subjects answer questions on the final test based on retrieval fluency, and that they do not carefully examine the source of the retrieved information. The goal of this paper is to provide support for this retrieval fluency account using a converging evidence approach. To that end, we examined (1) the effects of warning on RES, (2) confidence, and (3) response latencies.

Confidence and retrieval latency

Research suggests that metamemorial assessments may be influenced by the ease with which information comes to mind (Koriat, 1993; Koriat, Ma'ayan, & Nussinson, 2006). For example Nelson and Narens (1990) proposed that confidence in answers is in part determined by retrieval latency. Supporting this conclusion, they found a negative correlation between confidence judgment and response latency. That is, the faster the response, the higher the confidence in that response. Further, this relation held for both accurate and incorrect recall. In a task that involved answering general knowledge questions, Kelley and Lindsay (1993) manipulated retrieval fluency by priming subjects with correct or semantically related, but incorrect, answers prior to questioning. Similar to Nelson and Narens, they found that confidence was negatively correlated with latency. Kelley and Lindsay argued that pre-exposure to correct and to related but incorrect answers caused those answers to come to mind easily and quickly, and the ease with which those answers came to mind led to high confidence.

Research has also demonstrated that like prior exposure, post-event questioning and post-event reflection (i.e., mentally reviewing and evaluating one's previous responses) affected confidence in final answers. Specifically, Shaw (1996) demonstrated that repeated testing paired with reflection on those initial responses led to higher confidence ratings on a later, final test, and suggested that the question-reflection pairing increased retrieval fluency of those answers. Additionally, presentation of a narrative with information consistent or inconsistent with an originally witnessed event resulted in higher confidence on a final test than when a general narrative was presented (Bonham & González-Vallejo, 2009).

In the context of RES, because initial testing enhances encoding of details in the post-event narrative, it should also increase the retrieval fluency of these details. Therefore, we expected that initial testing would increase **confidence** for responses associated with information presented in the post-event narrative in Experiment 1, regardless of whether that information is correct or misleading. To provide additional, and perhaps more direct, support for the retrieval fluency hypothesis, we examined the latency of responses in a recognition test in Experiment 2. Response latency is considered a relatively direct measure of retrieval fluency (Benjamin, Bjork, & Schwartz, 1998). As such, we hypothesized that initial testing would lead to faster response times on the final test when subjects responded with details they learned from the post-event narrative (e.g., the misinformation).

Manipulating retrieval strategy via warning

The retrieval fluency hypothesis specifies that initial testing causes the misinformation to come to mind easily during the final test, which in turn leads subjects to prematurely terminate further recollection that is needed to recall the original target information (e.g., Jacoby, Bishara, Hessels, & Toth, 2005; Jacoby & Rhodes, 2006). To test this hypothesis, the present study examined whether subjects could be encouraged to engage in more effortful recollection and reduce inaccuracies by warning them about the veracity of the narrative. The effects of warning on eyewitness suggestibility have been investigated extensively. For example, Echterhoff, Hirst, and Hussy (2005b) (see also Chambers & Zaragoza, 2001b; Christiaansen & Ochalek, 1983; Eakin, Schreiber, & Sergent-Marshall, 2003; Greene, Flynn, & Loftus, 1982) found that warning subjects after misinformation exposure reduced the misinformation effect. In the context of RES, warning should encourage subjects to engage in more effortful recollection during retrieval (Starns, Lane, Alonzo, & Roussel, 2007), thereby reducing fluency-based responding. Thus, warning should reduce the influence of misinformation and its effect should be particularly pronounced after initial testing. This prediction is based on findings that testing can reduce interference (Szpunar, McDermott, & Roediger, 2008b) and enhance source memory (Chan & McDermott, 2007); therefore, providing subjects with a warning might allow the benefits of testing on source monitoring to surface. That is, when warned, initial testing might help, rather than hurt, subsequent eyewitness memory performance. With regard to response time measurements, providing a warning should reduce fluency-based responding for all subjects, which should result in an overall increase in response latencies. However, those who have taken an initial test would need to engage in more effortful recollection to override to prepotent, fluency-driven responses (i.e., the well-learned misinformation). As a result, when warned, the repeated testing subjects should produce longer response times than the single testing subjects.

Experiment 1

The goal of Experiment 1 was to examine whether the hypothesized increased retrieval fluency under RES

conditions could be minimized if subjects were warned about the validity of the post-event narrative. We hypothesized that warning would encourage subjects to engage in more controlled recollection, which should promote effective source discrimination. Additionally, Experiment 1 examined confidence associated with retrieved information. We expected to find higher confidence ratings in conditions where retrieval fluency was heightened. Specifically, these conditions included consistent and misleading information presented in the post-event narrative for subjects who had taken an initial test.

Method

Participants

Sixty-eight undergraduate students from Colby College and 12 undergraduates from Tufts University participated in this experiment for course research credit.

Materials and procedure

The experimental design was a 2 (warning: no warning, warning) × 2 (testing: single, repeated) × 3 (item type: consistent, control, misleading) mixed design. Warning and testing were manipulated between-subjects, whereas item type was manipulated within-subjects. Subjects first viewed a ∼40 min episode of the television program "24" (the witnessed event). We used the first episode of the first season of "24" as the witnessed event material. The audio narrative was created by modifying the episode guide provided by Fox television at www.fox.com/24. No subjects had seen this video before.

After viewing the witnessed event video, subjects in the **repeated testing condition** took an immediate recall test on 24 details of the video (e.g., Question: What did the terrorist use to knock out the flight attendant? Answer [not provided to subjects]: A hypodermic syringe), whereas subjects in the **single test condition** played Tetris (a computerized falling-rock puzzle game) for the same amount of time (12 min). During the cued recall test, subjects were told to answer every question (by typing their responses into the computer) and then indicate their confidence ranging from 0% to 100%. They were instructed to give a confidence rating of zero for guesses. No corrective feedback was provided. All subjects then completed a brief demographic questionnaire, a synonym and antonym vocabulary test (Salthouse, 1993) and computerized Operation Span (OSPAN, Kane & Engle, 2003) as distractor tasks. This distractor phase lasted approximately 20 min.

Following the distractor tasks, subjects listened to an 8 min audio narrative that described the video. Subjects in the **no-warning** condition were told that the narrative was a recap of the video (the experimenter did not warn subjects about the veracity of the narrative). After the narrative was played and before the final test, subjects in the **warning** condition were told: "You will have to answer questions regarding the episode you previously watched. We just played a narrative of that episode; however, we are uncertain as to the source of the narrative. Therefore, we were unable to verify the accuracy of the narrative. As such, base your answer only on what you saw in the episode, and not on what you heard in the narrative".

Of the 24 details targeted by the initial test, eight of them were presented correctly in the narrative (*consistent.* e.g., [the terrorist] knocks the flight attendant unconscious with a *hypodermic syringe*), eight were not mentioned in the narrative (*control.* e.g., [the terrorist] knocks the flight attendant unconscious), and eight were changed in the narrative (*misleading.* e.g., [the terrorist] knocks the flight attendant unconscious with a *chloroform rag*). The misleading information always involved replacing a specific item with a plausible alternative. Each critical detail appeared only once in the narrative and whether the detail was consistent, control, or misleading was counterbalanced across subjects. Both focal and non-focal details were modified. The final test was identical to the initial test and subjects were told to report the information presented in the video.

Results

Cued recall

Unless otherwise stated, *p*-values are less than .05. During the initial recall test, .61 of subjects' responses were accurate and .06 matched the misinformation spontaneously (i.e., baserate false recall). We now examine the accurate recall probability on the final test.

The top half of Table 1 presents the accurate recall probabilities in the final test. Separate 3 (item type: consistent, control, misleading) × 2 (testing condition: single, repeated) analysis of variance (ANOVAs) examined the effects of item type and testing on final accurate recall for each warning group. There was an interaction between item type and testing condition for subjects in the **no-warning** condition, $F(2, 76) = 8.46$, $MSe = .02$, such that initial testing increased accurate recall for the consistent items, $t(38) = 2.20$, $d = .71$, had no influence on the control items, $t < 1$, but decreased accurate recall for the misleading items, $t(38) = 2.99$, $d = .96$, (i.e., RES). These data suggest that subjects were particularly likely to recall a detail provided by the post-event narrative after they have taken an initial test, regardless of whether that detail was consistent with or contradictory to the original event. This finding is consistent with the possibility that subjects responded

Table 1
Average accurate recall (for Experiment 1) and hit rates (for Experiment 2) on the final test as a function of test condition, warning, and item type (standard deviations are in parentheses).

	Consistent	Control	Misleading
Experiment 1 (cued recall)			
No warning			
Single testing	.73 (.20)	.60 (.19)	.44 (.14)
Repeated testing	.85 (.13)	.58 (.11)	.28 (.19)
Warning			
Single testing	.68 (.15)	.63 (.16)	.58 (.14)
Repeated testing	.77 (.13)	.72 (.17)	.69 (.15)
Experiment 2 (recognition)			
No warning			
Single testing	.86 (.14)	.77 (.14)	.62 (.22)
Repeated testing	.86 (.12)	.78 (.15)	.41 (.29)
Warning			
Single testing	.83 (.13)	.74 (.18)	.59 (.23)
Repeated testing	.84 (.18)	.76 (.18)	.67 (.17)

based on recency or retrieval fluency. Critically, the interaction between item type and testing condition was eliminated for subjects in the **warning** condition, $F < 1$. Instead, a main effect of testing was found, $F(1, 38) = 8.56$, $MSe = .03$, such that initial testing boosted recall performance on the final test ($M = .73$ for repeated testing and $M = .63$ for single testing), and this testing benefit occurred regardless of whether an item was contradicted by later misinformation. This finding is consistent with the notion that warning reduced fluency-driven responding that contributes to RES.

To examine the effects of initial testing on recall probability of misinformation, a 2 (testing condition: single, repeated) × 2 (warning: no warning, warning) ANOVA was performed on misinformation production on the final test. There was a main effect of testing condition, $F(1, 76) = 7.67$, $MSe = .02$, a main effect of warning, $F(1, 76) = 31.88$, $MSe = .02$, and a significant interaction between them, $F(1, 76) = 6.74$, $MSe = .02$. As Table 2 illustrates, in the **no-warning** condition, repeated testing increased misinformation production compared to a single test (i.e., an .18 RES effect, $t(38) = 3.78$, $d = 1.16$). However, the RES effect was virtually eliminated when subjects were **warned**, $t < 1$! This finding represents a boundary condition for RES. Indeed, when equipped with a warning, subjects were able to reap the benefits of repeated retrieval without falling prey to RES.

Confidence

Although subjects provided confidence ratings for both the initial and final test, we only analyzed data from the final test (see Table 3). Again, we conducted ANOVAs for subjects in the no-warning and the warning conditions separately. The 3 (item type: consistent, control, misleading) × 2 (testing condition: single, repeated) ANOVA in the **no-warning** condition found an interaction between item type and warning, $F(2, 76) = 3.66$, $MSe = 137.3$. Specifically, initial testing increased confidence for the consistent and misleading questions during the final test, both $t_s > 3.72$, $d_s > 1.18$, but no difference was observed for the control questions, $t = 1$. The 3 (item type: consistent, control, misleading) × 2 (testing condition: single, repeated) ANOVA in the **warning** condition also found a significant interaction, $F(2, 76) = 4.09$, $MSe = 112.6$. However, this interaction was driven by the increase in confidence judgments associated with control trials following repeated testing, $t(38) = 2.84$, $d = .88$. In contrast, testing had no ef-

Table 2
Average misinformation production (for Experiment 1) and false recognition of misinformation (for Experiment 2) associated with misleading trials (standard deviations in parentheses).

No Warning Single testing	Repeated testing	Warning Single testing	Repeated testing
Experiment 1 (cued recall)			
.30 (.14)	.48 (.17)	.19 (.16)	.20 (.15)
Experiment 2 (recognition)			
.31 (.21)	.53 (.29)	.31 (.27)	.24 (.18)

Table 3
Average confidence judgments (ranging from 0–100) associated with the final test as a function of warning, test condition and item type (standard deviations in parentheses).

	Consistent	Control	Misleading
Experiment 1 (cued recall)			
No warning			
Single testing	72 (17.94)	58 (16.40)	63 (16.66)
Repeated testing	90 (10.86)	64 (16.91)	80 (11.68)
Warning			
Single testing	75 (11.22)	67 (14.83)	70 (14.76)
Repeated testing	73 (14.09)	78 (9.62)	71 (15.64)
Experiment 2 (recognition)			
No warning			
Single testing	79 (15.76)	68 (18.51)	76 (15.35)
Repeated testing	94 (7.57)	75 (14.87)	84 (10.97)
Warning			
Single testing	84 (10.04)	81 (15.22)	81 (10.82)
Repeated testing	83 (15.05)	80 (10.82)	66 (12.53)

fect on confidence judgments associated with consistent and misleading trials, both $t_s < 1$.[2]

A 2 (testing condition: single, repeated) × 2 (warning: no warning, warning) ANOVA was conducted on confidence judgments associated with misinformation production. The interaction between testing condition and warning was significant, $F(1, 58) = 8.72$, $MSe = 570.79$. As Table 4 illustrates, without a warning, initial testing dramatically increased confidence in misinformation production (a 33-point increase), $t = 3.84$, $d = 1.34$. Remarkably, when subjects were warned, this inflation in confidence was eliminated, $t < 1$.

Discussion for Experiment 1

The primary goal of Experiment 1 was to explore a retrieval fluency explanation for RES by using a warning manipulation and examining confidence judgments. Based on this explanation, warning should reduce fluency-based responding and encourage subjects to engage in more effortful recollection. Consistent with this prediction, warning enhanced recall accuracy for both the single testing and repeated testing groups, but this advantage was particularly pronounced for the repeated testing group. Specifically, in the single-testing condition, warning improved recall performance for the misleading trial (a 14% increase in accurate recall and 11% reduction in misinformation production). This finding is consistent with numerous studies that have demonstrated that the misinformation effect can be mitigated by warning (e.g.,

---

[2] Two 3 (item type: consistent, control, misleading) × 2 (testing: single, repeated) ANOVAs were performed on confidence judgments conditionalized on correct retrieval for both warning conditions. We chose not to include these analyzes as confidence associated with correct responding did not differ as a function of item, warning, or testing. Analyzes on confidence judgments associated with incorrect responses were also not included for consistent and control items, because incorrect responding in these conditions was infrequent. Thus, the analyzes on confidence judgments in association with these trials include confidence associated with both correct and incorrect responding.

Table 4
Confidence (ranging from 0 to 100) associated with misleading trials (standard deviations in parentheses).

	Correct answer	Misinformation
Experiment 1 (cued recall)		
No warning		
Single testing	83 (19.02)	58 (32.92)
Repeated testing	85 (14.18)	92 (11.56)
Warning		
Single testing	85 (17.72)	63 (19.66)
Repeated testing	85 (16.06)	61 (24.96)
Experiment 2 (recognition)		
No warning		
Single testing	87 (14.45)	71 (25.01)
Repeated testing	82 (31.17)	73 (26.00)
Warning		
Single testing	90 (12.85)	81 (17.57)
Repeated testing	86 (12.53)	43 (24.57)

Chambers & Zaragoza, 2001a; Echterhoff, Hirst, & Hussy, 2005a; Lindsay, 1990; Wright, 1993). However, for subjects in the repeated testing condition, warning increased accurate recall for both the control and misleading items. The improvement for the misleading questions was particularly dramatic (a 41% increase in accurate recall and 28% reduction in misinformation production). From the perspective of RES, where the comparison of interest is between the single testing and repeated testing condition, warning represents a powerful boundary to retrieval-enhanced suggestibility. When equipped with a warning, subjects were able to increase accurate recall by 11% while keeping misinformation production the same with repeated testing.

In addition to accuracy, confidence judgments can provide further support for the retrieval fluency hypothesis. Previous research suggests that retrieval fluency may serve as a basis for confidence judgments (Stretch & Wixted, 1998; Van Zandt, 2000). As such, we expected to find higher confidence ratings in conditions where retrieval fluency was heightened. Specifically, we expected high confidence when consistent information was presented in the narrative and reported on the final test, as well as high confidence when misleading information was presented in the narrative and reported on the final test. Consistent with these predictions, when subjects were not warned, those in the repeated testing condition gave higher confidence ratings on the final test for consistent and misleading items as compared to subjects in the single-testing condition. When given a warning, subjects in the repeated testing group were less likely to be influenced by retrieval fluency, and this test-induced inflation in confidence disappeared.

Experiment 2

In Experiment 1, we used warning and confidence to support a retrieval fluency explanation for the RES effect. Similar to Chan et al. (2009), when subjects were not warned but had received a test prior to misinformation, they were more likely to produce misinformation on the final test than subjects who had not received the initial test. While various mechanisms can account for this finding (cf.

Chan & Langley, in press; Chan et al., 2009), Experiment 1 supports a retrieval fluency explanation for two reasons. First, when subjects in the repeated testing condition were warned, accuracy for misleading items significantly increased. This finding provides indirect support for the retrieval fluency explanation. That is, subjects were able to override the influence of information in the narrative by engaging more careful, effortful recollection. Such recollections were more successful with repeated testing because retrieval practice has been demonstrated to enhance recollection and source memory (Chan & McDermott, 2007). Second, when unwarned, repeated testing increased confidence judgments for responses associated with consistent and misleading questions, but not for control questions. These results suggest that initial testing enhanced encoding of details in the post-event narrative. Upon final test, those details were quickly and easily retrieved. Speed and ease of retrieval has consistently been demonstrated to influence confidence, as these cues serve as an indicator for memorial accuracy. In the present experiment, those cues were misleading.

Experiment 2 was designed to provide additional support for the retrieval fluency explanation of RES. In addition to confidence, here we also examined response latency – a more direct measurement of retrieval fluency. We operationalize retrieval fluency as the speed with which information is accessed and reported from memory (e.g., Benjamin et al., 1998). To that end, we employed a four alternative forced choice test and examined response latencies associated with recognition. Additionally, we were interested in whether the RES finding would manifest in recognition (see Ayers & Reder, 1998; Loftus, 2005, for reviews of the misinformation effect in recognition). In support of the retrieval fluency hypothesis, we expected that subjects in the repeated testing condition would choose misleading and consistent information on the final test more quickly than subjects in the single-testing condition. Further, warning should have a greater effect on response latencies for the repeated testing group than for the single testing group, and its effects would be particularly pronounced for the misleading questions. For these misleading trials, because subjects would have access to two conflicting sources of information (one from the originally witnessed event and one from the post-event narrative), providing a warning should encourage subjects to engage more effortful recollective processes to override the fluently-retrieved misinformation, thereby leading to slower response times.

Method

Participants

Sixty-six undergraduate students from Tufts University participated in this experiment for course research credit or for payment of $15.

Materials and procedure

All experimental protocols were the same between Experiment 1 and Experiment 2 except that the memory test was changed from cued recall to recognition (for both the initial and final tests). Each test question featured four

alternatives. One alternative was the correct item. One was the misleading lure. The remaining two were plausible lures. Pilot testing ensured that incorrect items (including the misleading item) were similarly chosen in the absence of misleading post-event information (see Bulevich, 2007 for lure construction). After selecting each answer, subjects indicated how confident they were in each answer. Subjects were instructed to answer the questions as quickly and accurately as possible. They were told that measures of response latency were being collected.

Results

Recognition

During the initial recall test, .76 of subjects' responses were accurate and .08 matched the misinformation spontaneously. We now examine the accurate recall probability on the final test. A 3 (item type: consistent, control, misleading) × 2 (testing condition: single, repeated) ANOVA was conducted for each warning condition separately (see the bottom half of Table 1 for the means). An interaction between item type and testing condition was found for the **no-warning** condition, $F(2, 64) = 4.18$, $MSe = .03$. That is, comparing to single testing, repeated testing significantly reduced the hit rate of the misleading trials (a .21 reduction, $t(32) = 2.34$, $d = .82$), but it did not affect the hit rate of the consistent and control trials, both $t_s < 1$. Similar to Experiment 1, this interaction was eliminated for subjects in the **warning** condition, $F < 1$, such that warning effectively removed the response bias (i.e., responding based on retrieval fluency) following initial testing.

To examine the effects of repeated testing on susceptibility to misinformation, a 2 (testing condition: single, repeated) × 2 (warning: no warning, strong warning) ANOVA was performed on **false recognition of misinformation** (see bottom half of Table 2). A crossover interaction was observed, $F(1, 62) = 6.05$, $MSe = .06$ (see Table 2 for means), such that repeated testing produced a powerful ($M = .22$) RES effect without a warning, $t(32) = 2.62$, $d = .87$, but this RES effect was eliminated (a non-significant 7% reduction) with a warning, $t < 1$. Alternatively, subjects in the repeated testing condition were far more likely to benefit from the warning and reduce false recognition of misinformation (.53 without warning vs. 24 with warning), $t(29) = 3.32$, $d = 1.20$, than subjects in the single-testing condition (no difference).

Confidence

Confidence ratings for answers on the final test were analyzed (see bottom half of Table 3). A 3 (item type: consistent, control, misleading) × 2 (testing condition: single, repeated) ANOVA was conducted for data in each warning condition separately. A significant interaction between item type and testing condition was not found when data in the **no-warning** condition were examined, $F = 1.73$; however, a main effect of item type was found, $F(2, 64) = 12.21$, $MSe = 156.20$. Confidence was highest for consistent items overall [consistent-misleading: $t(33) = 2.48$, $d = .45$; consistent-control: $t(33) = 4.58$, $d = .95$]. Confidence was also higher for misleading items as compared to control items, $t(33) = 2.62$, $d = .56$. A main effect of item type

was found when data in the **warning** condition were analyzed, $F(2, 60) = 6.05$, $MSe = 129.67$. The overall confidence ratings were higher on consistent and control trials than on misleading trials [consistent-misleading: $t(31) = 2.78$, $d = .67$; control-misleading $t(31) = 1.91$, $p = .07$, $d = .45$]. A significant interaction was also found, $F(2, 60) = 4.27$, $MSe = 129.67$. That is, relative to single testing, repeated testing reduced confidence of accurate recall on misleading trials, $t(30) = 3.74$, $d = 1.28$, but not on the consistent or control trials, all $t_s < 1$. No other item-type comparisons were significant.

A 2 (testing condition: single, repeated) × 2 (warning: no warning, warning) ANOVA conducted on confidence judgments associated with **false recognition of misinformation** (see the bottom half of Table 4) yielded a significant interaction between testing condition and warning, $F(1, 61) = 11.69$, $MSe = 545.92$. Specifically, repeated testing had no effect on confidence associated with false recognition when subjects were not warned, $t < 1$, but it substantially lowered the confidence for false recognition when subjects were warned, $t(30) = 5.10$, $d = 1.78$. Remarkably, false recognition confidence dropped by nearly 40% when the repeated testing condition was compared with those in the single-testing condition!

Response latencies

Response latencies associated with recognition allowed us to further examine the retrieval fluency explanation for the RES. We hypothesized that initial testing would strengthen encoding of the details in the post-event narrative, which would then affect the speed with which these details were accessed at final test. Specifically, we expected faster response times in the repeated testing condition than the single-testing condition. In addition, we expected that warning would slow down responses.

The response latency data in the final test are presented in Fig. 1. A 3 (item type: consistent, control, misleading) × 2 (testing condition: single, repeated) ANOVA in the **no-warning** condition (see Fig. 1a) found a main effect of item type, $F(2, 64) = 17.07$, $p < .001$, and a main effect of testing, $F(1, 32) = 56.67$, $p < .001$. Subjects responded more quickly on both the misleading trials, $t(33) = 4.07$, $d = .79$, and the consistent trials, $t(33) = 4.99$, $d = .76$ than they did on the control trials. In addition, subjects responded more quickly in the repeated testing condition ($M = 4340.61$ ms) as compared to the single-testing condition
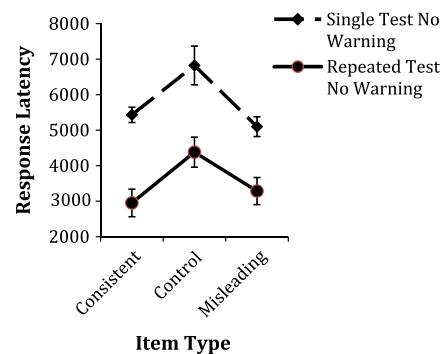

Fig. 1a. Response latencies as a function of item type and testing for the no-warning condition in Experiment 2 (error bars show standard error).

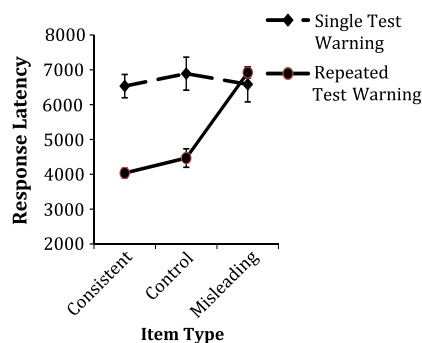**Fig. 1b.** Response latencies as a function of item type and testing for the warning condition in Experiment 2 (error bars show standard error).

Table 5
Mean response latencies (in ms) associated with recognition decisions for misleading trials in Experiment 2 (standard errors in parentheses).

	Correct answer	Misinformation
No warning		
Single testing	5743 (345)	7023 (713)
Repeated testing	3866 (244)	3492 (107)
Warning		
Single testing	5665 (270)	7389 (604)
Repeated testing	3811 (361)	6217 (470)

($M = 6227.21$ ms). Taken together, these data provide additional support that initial testing influenced the ease with which information from the post-event narrative came to mind on the final test.

A 3 (item type: consistent, control, misleading) × 2 (testing condition: single, repeated) ANOVA in the **warning** condition (see Fig. 1b) found a significant interaction, $F(2, 60) = 14.45$, $p < .001$. Unlike the no-warning condition, subjects in the repeated testing condition actually took significantly **longer** to make recognition decisions on misleading trials as compared to consistent, $t(13) = 6.37$, $d = 1.80$, and control trials, $t(13) = 6.01$, $d = 1.34$; however, subjects in the single-testing condition did not demonstrate a difference in response latencies as a function of item type, all $t_s < 1$. This pattern suggests that following a warning, subjects in the repeated testing condition (but not those in the single-testing condition) might have noticed the conflicting nature of the misinformation, thus slowing their responses for these questions relative to the other questions (i.e., the consistent and control questions).

Finally, we examined response latencies on **misleading trials** as a function of response type (correct vs. misinformation). A 2 (response type: correct, misinformation) × 2 (testing condition: single, repeated) × 2 (warning: no warning, warning) ANOVA found main effects of response type, $F(1, 60) = 31.60$, $p < .001$, and testing condition, $F(1, 60) = 30.60$, $p < .001$. The 3-way interaction was also significant, $F(1, 60) = 6.93$, $p = .01$. Planned comparisons were conducted to decompose this interaction. As Table 5 illustrates, when the repeated testing subjects were not warned, they responded with similar speed for correct and false recognition, $t(15) = 1.32$, $p = .19$. However, when the repeated testing subjects were warned, they slowed down considerably when they selected the misinformation as compared to when they selected the correct alternative, $t(13) = 5.34$, $d = 1.59$. In contrast, when subjects in the single testing group selected the misinformation, their response times were always longer than when they were correct, regardless of warning condition, both $t_s > 2.40$, $d_s > .56$. Taken together, these results are consistent with the idea that when subjects were not warned, repeated testing caused the misinformation to come to mind easily and subjects responded with this misinformation without carefully evaluating its source. However, once warned, the repeated testing subjects slowed down to avoid reporting the fluently-retrieved misinformation, and their recognition accuracy rose accordingly.

Discussion for Experiment 2

Results from response latencies provided additional support for the retrieval fluency explanation of RES. According to this explanation, enhanced suggestibility may in part be due to changes in retrieval fluency that result from the combination of the initial test and narrative presentation. Thus, we expected that when subjects were not warned, repeated testing would shorten response latencies during the final test (relative to single testing). This prediction was panned out in the data. Most importantly, we expected that warning would show its biggest influence on response latency when subjects must override the prepotent, fluency-driven misinformation. This was manifested as a dramatic slow down for the repeated testing subjects when they answered the misleading questions (compare no-warning to warning). Specifically, in cases where misinformation was selected, it took subjects in the repeated testing condition significantly longer to make this decision after having received a warning. Warning did not have an impact on response latencies associated with false recognition of misinformation for subjects in the single-testing condition.

General discussion

The goal of this study was to explore a retrieval fluency hypothesis for the RES effect. We hypothesized that prior retrieval influenced narrative processing such that the consistent and misleading information "popped out" and were more easily accessed on the final test. For subjects in the repeated testing condition, the misinformation may be similar to hard-to-inhibit prepotent responses (e.g., Hasher & Zacks, 1988). This hypothesis states that RES is driven partly by retrieval fluency and can be overridden by source-specifying recollective processes. Consistent with this notion, when warned, repeated testing increased cued recall and recognition accuracy compared to single testing. Moreover, warning (compared to no warning) led to longer latencies associated with misleading trials and a reduction in confidence on those trials, especially after initial testing.

Warning and testing increases memory accuracy

RES is a puzzling, yet powerful, phenomenon. When subjects are not warned about the veracity of the post-event

information, those who received repeated tests demonstrated more errors on the final test of memory, were more likely to produce or select misleading post-event information on the final test, and had higher confidence in these incorrect responses than subjects in the single-testing condition. However, when given a warning, the repeated testing advantage was observed, such that subjects in the repeated testing condition demonstrated overall better accuracy on the final test than subjects in the single-testing condition. That is, testing produced **opposite effects** on eyewitness suggestibility as a function of warning. These results suggest that unwarned subjects in the repeated testing condition responded with the most fluently retrieved item, but that response bias was mitigated by a warning.

Initial testing can enhance retention of originally learned material (Roediger & Karpicke, 2006; Tulving, 1967) and improve learning of new information (Szpunar, McDermott, & Roediger, 2008a; Tulving & Watkins, 1974). In the current context, testing should therefore strengthen memories for both the original witnessed details and the misleading details. The RES finding suggests that subjects are responding based on recency or retrieval fluency, and they are unlikely to demonstrate the testing effect without an intervention that directs them to carefully evaluate the information that comes to mind. Response latencies associated with recognition in Experiment 2 provide additional support for this argument. Specifically, subjects in the repeated testing condition who were not warned responded more quickly than all other groups of subjects on the consistent and misleading questions. Even on a recognition test, where the correct option was present, the bias developed by the test-narrative combination could not be successfully overcome without an explicit warning. However, when provided with a warning, subjects took time to evaluate the source of the retrieved information, and this effortful process led to increased accuracy.

We believe that warning allowed the testing advantage to be revealed because it reduced fluency-based responding in the current paradigm. This reduction may stem partly from an effortful retrieval process where subjects attempt to access multiple potential targets and evaluate the perceptual and contextual cues associated with those targets (Johnson, Hashtroudi, & Lindsay, 1993). This argument rests on the assumption that subjects in the repeated testing condition are able to access originally learned material, and is contrary to the reconsolidation argument proposed by Chan et al. (2009). Specifically, in Chan et al., we suggested that subjects in the repeated testing condition reactivated memories of the witnessed event during the initial test. Those memories then became particularly susceptible to interference. Results from a modified-modified free recall (MMFR) test, in which subjects in the repeated testing group continued to demonstrate a misinformation effect, supported this hypothesis. The present study clearly demonstrates that people can access the originally learned material under certain conditions (i.e., warning).

Retrieval fluency affects confidence

The high confidence judgments that accompanied responses on misleading and consistent trials provide additional support for the retrieval fluency explanation of RES. Research has consistently demonstrated that the ease with which information comes to mind serves as an indicator for confidence (Kelley & Lindsay, 1993; Nelson & Narens, 1990; Shaw, 1996). Consistent with the retrieval fluency hypothesis, the unwarned subjects in the repeated testing condition were extremely confident in their responses on both misleading and consistent trials in both experiments. The use of retrieval fluency as a cue for confidence often does yield good calibration between reported confidence and response accuracy. Memories that are easily retrieved are often accurate, so their accompanying high confidence is usually appropriate (Mandler & Boeck, 1974; Perfect & Hollins, 1999; Tulving & Thomson, 1971). This results in a robust relationship between confidence and memory. However, in the case of RES, retrieval-fluency is a poor indictor of accuracy, and it negatively affects the diagnosticity of confidence.

Retrieval-fluency is only one cue that has been shown to influence confidence. Additionally, multiple recollection attempts, memory vividness, access to corroborating detail, (Perfect & Hollins, 1999), as well as the completeness and amount of information retrieved (Brewer, Sampaio, & Barlow, 2005; Koriat, Lichtenstein, & Fischhoff, 1980), all have been shown to influence confidence. Just as retrieval fluency has been shown to be an unreliable index of accuracy, under certain circumstances these other cues also can inappropriately inform confidence in responses. For example, Brewer et al. demonstrated that when subjects felt that they had completely recalled a sentence, they indicated high confidence in their memory, even though the surface structure of the original sentence was incorrect. In the present study, responses about which a person was highly confident tended to be items that were quickly retrieved. Because repeated testing increased the speed with which recognition decisions were made on final test, the confidence in these responses was likely influenced by their response latencies. In the case of misleading trials, these quickly-accessed memories led to confidently-held false memories.

Conclusions

In the present study, we have demonstrated the consequences of initial testing on eyewitness suggestibility. When a witness is initially "tested," both the information from the retrieval cue and the information retrieved become particularly accessible. When new but related information is presented, this new information captures the attention of the witness and is better learned. Thus, information from the narrative, which has been differentially processed, becomes highly accessible. When subjects are not warned, they retrieve this information very fluently, leading to RES. However, when subjects are warned, they are more likely to engage in more effortful, controlled processes and moderate the powerful cue of retrieval fluency, and such recollection is particularly effective after initial testing (Chan & McDermott, 2007).

References

Ayers, M. S., & Reder, L. M. (1998). A theoretical review of the misinformation effect: Predictions from an activation-based memory model. *Psychonomic Bulletin & Review, 5*, 1–21.

Baddeley, A. D. (1982). Domains of recollection. *Psychological Review, 89*, 708–729.

Baddeley, A. D. (1982b). Domains of recollection. *Psychological Review, 89*(6), 708–729. <http://www.apa.org/journals/rev.html>.

Benjamin, A. S., Bjork, R. A., & Schwartz, B. L. (1998). The mismeasure of memory: When retrieval fluency is misleading as a metamnemonic index. *Journal of Experimental Psychology: General, 127*(1), 55–68.

Bonham, A. J., & González-Vallejo, C. (2009). Assessment of calibration for reconstructed eye-witness memories. *Acta Psychologica, 131*(1), 34–52.

Brewer, W. F., Sampaio, C., & Barlow, M. (2005). Confidence and accuracy in the recall of deceptive and nondeceptive sentences. *Journal of Memory and Language, 52*(4), 618–627.

Bulevich, J. B. (2007). Can controlled processes improve performance in the misinformation paradigm? *Dissertation Abstracts International: Section B: The Sciences and Engineering, 68*(6-B).

Chambers, K. L., & Zaragoza, M. S. (2001a). Intended and unintended effects of explicit warnings on eyewitness suggestibility: Evidence from source identification tests. *Memory & Cognition, 29*(8), 1120–1129.

Chambers, K. L., & Zaragoza, M. S. (2001b). Intended and unintended effects of explicit warnings on eyewitness suggestibility: Evidence from source identification tests. *Memory & Cognition, 29*, 1120–1129.

Chan, J. C. K., Langley, M. (in press). Paradoxical effects of testing: Retrieval enhances both accurate recall and suggestibility in eyewitnesses. Journal of Experimental Psychology: Learning, Memory, and Cognition.

Chan, J. C. K., & McDermott, K. B. (2007). The testing effect in recognition memory: A dual process account. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*(2), 431–437.

Chan, J. C. K., Thomas, A. K., & Bulevich, J. B. (2009). Recalling a witnessed event increases eyewitness suggestibility: The reversed testing effect. *Psychological Science, 20*(1), 66–73.

Christiaansen, R. E., & Ochalek, K. (1983). Editing misleading information from memory: Evidence for the coexistence of original and post-event information. *Memory & Cognition, 11*, 467–475.

Eakin, D. K., Schreiber, T. A., & Sergent-Marshall, S. (2003). Misinformation effects in eyewitness memory: The presence and absence of memory impairment as a function of warning and misinformation accessibility. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*, 813–825.

Echterhoff, G., Hirst, W., & Hussy, W. (2005a). How eyewitnesses resist misinformation: Social postwarnings and the monitoring of memory characteristics. *Memory & Cognition, 33*(5), 770–782.

Echterhoff, G., Hirst, W., & Hussy, W. (2005b). How eyewitnesses resist misinformation: Social postwarnings and the monitoring of memory characteristics. *Memory & Cognition, 33*, 770–782.

Greene, E., Flynn, M. S., & Loftus, E. F. (1982). Inducing resistance to misleading information. *Journal of Verbal Learning and Verbal Behavior, 21*, 207–219.

Hasher, L., & Zacks, R. T. (1988). *Working memory, comprehension, and aging: A review and a new view.* San Diego, CA: Academic Press.

Jacoby, L. L., Bishara, A. J., Hessels, S., & Toth, J. P. (2005). Aging, Subjective Experience, and Cognitive Control: Dramatic False Remembering by Older Adults. *Journal of Experimental Psychology: General, 134*(2), 131–148.

Jacoby, L. L., & Dallas, M. (1981). On the relationship between autobiographal memory and perceptual learning. *Journal of Experimental Psychology: General, 110*, 306–340.

Jacoby, L. L., Kelley, C. M., & Dywan, J. (1989a). *Memory attributions.* Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc..

Jacoby, L. L., Kelley, C. M., & Dywan, J. (1989b). Memory attributions. In H. L. Roediger & F. I. M. Craik (Eds.), *Varieties of memory and consciousness: Essays in honour of Endel Tulving* (pp. 391–422). Hillsdale, NJ: Erlbaum.

Jacoby, L. L., & Rhodes, M. G. (2006). False Remembering in the Aged. *Current Directions in Psychological Science, 15*(2), 49–53.

Johnson, M. K., Hashtroudi, S., & Lindsay, D. (1993). Source monitoring. *Psychological Bulletin, 114*(1), 3–28.

Kane, M. J., & Engle, R. W. (2003). Working-memory capacity and the control of attention: The contributions of goal neglect, response competition, and task set to Stroop interference. *Journal of Experimental Psychology: General, 132*, 47–70.

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

Koriat, A. (1993). How do we know that we know? The accessibility model of the feeling of knowing. *Psychological Review, 100*(4), 609–639.

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory, 6*(2), 107–118.

Koriat, A., Ma'ayan, H., & Nussinson, R. (2006). The intricate relationships between monitoring and control in metacognition: Lessons for the cause-and-effect relation between subjective experience and behavior. *Journal of Experimental Psychology: General, 135*(1), 36–69.

Lindsay, D. S. (1990). Misleading suggestions can impair eyewitnesses' ability to remember event details. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*(6), 1077–1083.

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

Loftus, E. F., Miller, D. G., & Burns, H. J. (1978). Semantic integration of verbal information into a visual memory. *Journal of Experimental Psychology: Human Learning and Memory, 4*(1), 19–31.

Mandler, G., & Boeck, W. J. (1974). Retrieval processes in recognition. *Memory & Cognition, 2*(4), 613–615.

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. H. Bower (Ed.). *The psychology and learning and motivation* (Vol. 26, pp. 125–173). New York: Academic Press.

Perfect, T. J., & Hollins, T. S. (1999). Feeling-of-knowing judgments do not predict subsequent recognition performance for eyewitness memory. *Journal of Experimental Psychology: Applied, 5*(3), 250–264.

Robbins, D., & Irvin, J. R. (1976). The priority effect: Test effects on negative transfer and control lists. *Bulletin of the Psychonomic Society, 8*, 167–168.

Roediger, H. L., III, & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17*(3), 249–255.

Salthouse, T. A. (1993). Influence of working memory on adult age differences in matrix reasoning. *British Journal of Psychology, 84*(2), 171–199.

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

Starns, J. J., Lane, S. M., Alonzo, J. D., & Roussel, C. C. (2007). Metamnemonic control over the discriminability of memory evidence: A signal detection analysis of warning effects in the associative list paradigm. *Journal of Memory and Language, 56*, 592–697.

Stretch, V., & Wixted, J. T. (1998). Decision rules for recognition memory confidence judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*(6), 1397–1410.

Szpunar, K. K., McDermott, K. B., & Roediger, H. L. (2008a). Testing during study insulates against the build-up of proactive interference. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*, 1392–1399.

Szpunar, K. K., McDermott, K. B., & Roediger, H. L. III, (2008b). Testing during study insulates against the buildup of proactive interference. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*(6), 1392–1399.

Tulving, E. (1967). The effects of presentation and recall of material in free-recall learning. *Journal of Verbal Learning & Verbal Behavior, 6*(2), 175–184.

Tulving, E., & Thomson, D. M. (1971). Retrieval processes in recognition memory: Effects of associative context. *Journal of Experimental Psychology, 87*(1), 116–124.

Tulving, E., & Watkins, M. J. (1974). On negative transfer: Effects of testing one list on the recall of another. *Journal of Verbal Learning and Verbal Behavior, 13*, 181–193.

Van Zandt, T. (2000). ROC curves and confidence judgments in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*(3), 582–600.

Wright, D. B. (1993). Misinformation and warnings in eyewitness testimony: A new testing procedure to differentiate explanations. *Memory, 1*(2), 153–166.