Effects of Post-Warning Specificity on Memory Performance and Confidence in the Eyewitness Misinformation Paradigm

Philip A. Higham

University of Southampton, UK

Hartmut Blank

University of Portsmouth, UK

Karlos Luna

University of Minho, Portugal

In Press: Journal of Experimental Psychology: Applied

Author Note

Philip A. Higham, Department of Psychology, University of Southampton; Hartmut Blank, Department of Psychology, University of Portsmouth; Karlos Luna, Psychology Research Centre, University of Minho, Portugal.

Correspondence concerning this article should be addressed to either Philip A. Higham, Department of Psychology, University of Southampton, Highfield, Southampton, UK, SO17 1BJ or Hartmut Blank, University of Portsmouth, King Henry Building, King Henry I Street, Portsmouth, UK, PO1 2DY. E-mail: higham@soton.ac.uk or Hartmut.Blank@port.ac.uk, respectively. Thanks to Ryan Fitzgerald for helpful suggestions on this paper.

Abstract

The influence of post-event misinformation on memory is typically constrained by post-warnings (Blank & Launay, 2014), but little is known about the effectiveness of particular features of post-warning, such as their specificity. Experiment 1 compared two levels of post-warning specificity: A general post-warning just stated the presence of misinformation, whereas a specific post-warning identified the test items for which misinformation had been presented earlier. The specific post-warning, but not the general post-warning, eliminated both the misinformation effect and its deleterious impact on memory monitoring (using a classic two-alternative forced-choice recognition procedure). Experiment 2 ruled out an alternative interpretation of these findings and replicated this post-warning specificity pattern using a cued-recall test. In addition to the moderating influence of task representations on misinformation acceptance, we also observed two unexpected facilitative effects on event memory caused by misinformation. Misinformation facilitated event memory during narrative encoding if discrepancies between the event and the narrative were detected (Experiment 1) and during retrieval if a specific post-warning was combined with cued recall (Experiment 2). We interpret the facilitative effect of discrepancy detection within Jacoby, Wahlheim and Kelley's (2015) recursive-remindings framework on noticing and recollecting change.

Keywords: misinformation; post-warnings; suggestibility; discrepancy detection; recursive reminding

Public Significance Statement

We investigated whether warning people after-the-fact about the presence of misinformation affected their susceptibility to it. Such research is critical given the amount of misinformation people are exposed to in the form of "fake news." We found that specific warnings that highlighted particular questions associated with misinformation were particularly effective at helping people overcome misinformation and sometimes such warnings even facilitated their memory for the truth.

Effects of Post-Warning Specificity on Memory Performance and Monitoring in the Eyewitness Misinformation Paradigm

Many studies on eyewitness memory have shown that misleading information encoded after witnessing an event has a deleterious effect on memory reports (e.g., Blank, 1998; Echterhoff, Hirst, & Hussy, 2005; Higham, 1998; Higham, Luna, & Bloomfield, 2011; Lindsay, 1990; Lindsay & Johnson, 1989; Loftus, Miller, & Burns, 1978; Luna & Migueles, 2009; Wright, 1993; Zaragoza & Lane, 1994; Zaragoza & Mitchell, 1996; see Loftus, 2005 for a review). Loftus et al. (1978) introduced a three-stage paradigm for investigating the effect of misleading information on memory. As an example of this paradigm, an eyewitness might watch a videotape of a burglar stealing a wristwatch (*event*) and then read a misleading narrative summarizing the event in which it is stated, "the burglar stole a wallet" (*post-event misinformation*). A *misinformation effect* occurs when misled eyewitnesses are more likely than non-misled eyewitnesses to indicate on a final memory test for the event that they remember seeing a wallet being stolen in the videotape.

An important issue to address is whether people who mistakenly accept misinformation lack confidence in their decision, or whether they fully endorse it. The effect of misinformation on confidence is dependent on a number of factors, but several studies have suggested the latter. For example, Loftus, Donders, Hoffman, and Schooler (1989) found that misinformed participants responded as confidently to incorrect post-event details as they did to their memories of event details, leading them to claim that post-event misinformation created memories that are "quickly accessed and confidently held" (p. 607; see also Luna & Migueles, 2009). Henceforth, we refer to this pattern of impaired accuracy coupled with high confidence in endorsements of post-event details as the *signature pattern* of misinformation.

Warnings and Misinformation Effects

Unsurprisingly, memory theorists have investigated whether the effect of misinformation is moderated by warnings about the presence of misleading information. Previous warnings used in misinformation experiments have certainly been quite diverse, but they can be classified into two main groups: pre-warnings and post-warnings. Pre-warnings are given prior to the encoding of the post-event misinformation and research has generally shown that they are very effective at reducing misinformation effects (e.g., Dodd & Bradshaw, 1980; Greene, Flynn, & Loftus, 1982; V. L. Smith & Ellsworth, 1987), most likely because participants can attend to and identify the misinformation when it is first presented.

The evidence on post-warnings – typically issued right before the final memory test – is more mixed. In a recent meta-analysis, Blank and Launay (2014) established that, on average, post-warnings reduced the misinformation effect to less than half of its usual size; however, there was considerable variability across studies. In some studies, post-warnings completely eliminated misinformation effects (Blank, 1998; Christiaansen & Ochalek, 1983; Eakin, Schreiber, & Sergent-Marshall, 2003; Highhouse & Bottrill, 1995; Oeberst & Blank, 2012; Wright, 1993), whereas in other studies, robust misinformation effects persisted (e.g., Belli, Lindsay, Gales, & McCarthy, 1994; Frost, Ingraham, & Wilson, 2002; Greene et al., 1982; Higham, 1998; Higham et al., 2011; Lindsay, 1990, similar source condition).

Warning Specificity

Such variability of post-warning effects is not surprising given the heterogeneity of procedures used in different studies. Blank and Launay (2014) classified post-warnings along three dimensions, (1) their specificity (in terms of locating the misinformation) and the presence or absence of (2) social post-warning (i.e., discrediting the reliability of the source of the post-event information) and (3) "enlightenment" (a debriefing-like explanation of the

context and purpose of the previous deceptive introduction of misinformation; e.g., Oeberst & Blank, 2012). In this article, we draw attention to the specificity dimension, exploring a new aspect of post-warning specificity not included in Blank and Launay's (2014) analysis. Specifically, we examine the impact of a particular type of post-warning specificity – *item-specific post-warnings* about the presence of misinformation – on measures of memory performance and associated metacognition. We believe this type of post-warning to be considerably more effective than general post-warnings.

What do we mean by item-specific post-warnings precisely? In addition to noncritical filler questions, a standard memory test used in misinformation studies will typically include questions that probe memory for details that have been the target of misinformation (henceforth *misleading questions*) as well as questions probing memory for details not associated with misinformation (henceforth *control questions*). When a general post-warning (i.e., about the mere presence of misinformation) is provided along with such a test, participants still face uncertainty with respect to (1) how many misleading details had been presented and (2) the particular questions in the memory test the post-warning pertains to – and this uncertainty could lead to less effective memory search and monitoring strategies (see below). By contrast, item-specific post-warnings clearly identify test questions about items that had been the target of misinformation. Plainly speaking, item-specific post-warnings clearly distinguish "dangerous" (misleading) questions, that is, questions for which more elaborate search and monitoring is advisable, from "safe" (control) questions for which such caution is not necessarily required.

There are both theoretical and applied reasons to be interested in the effect of warning specificity. On the theoretical side, a long-standing explanation of the misinformation effect is "overwriting" or "destructive updating" (e.g., Loftus, 1979a, 1979b; Loftus & Loftus, 1980; Loftus et al., 1978). On this view, the original event memory is overwritten or

destroyed by misinformation. Although the original overwriting hypothesis has largely fallen out of favor over the years, largely due to the finding that original event memories are sometimes retrievable despite supposed overwriting (e.g., Bekerian & Bowers, 1983; McCloskey & Zaragoza, 1985), recently it has enjoyed a reappearance in the form of *reconsolidation-based memory impairment*. Chan and LaPaglia (2013) conducted a series of classical, three-stage misinformation experiments (i.e., event, post-event misinformation, memory test), except that half the participants were given a cued-recall pretest about the event details before receiving the misinformation whereas the other half were not. On a later true-false test that followed receipt of misinformation, they found that participants had impaired memory for original event details, but only if those memories had been earlier reactivated by the cued-recall pretest. They argued that reactivating memories (with the cued-recall test) produced a reconsolidation window during which the original memory must be restabilized. If misinformation is encoded during this window, this reconsolidation process is interrupted and can cause the original memory to be overwritten. They argued that their results "demonstrate that human declarative memory can be selectively rewritten during reconsolidation" (p. 9309; although see Rindal, DeFranco, Rich, & Zaragoza, 2016 for counterarguments to this claim).

If the overwriting hypothesis has any validity, then post-warnings, regardless of whether original event memories are reactivated and no matter how specific, should not influence retrieval of event details. Simply put, if the warning is given after misinformation has been encoded, it will be of no help in retrieving event details no matter how specific the warning is because the original event memory has gone. Thus, in terms of event memory, the destructive updating hypothesis predicts comparable performance between groups of

participants given general vs specific warnings.[1]

On the applied side, post-warning specificity is of particular relevance to the modern phenomena of "fake news" and "post-truth politics." We currently live in an era of rampant misinformation. Websites, social media, mainstream news, and even the current White House espouse facts and figures that have little to no basis in reality. The sheer prevalence of misinformation puts those who are interested in separating facts from fiction in an awkward position: What should be believed and what should be taken with a grain of salt? One tool that news consumers have available to them is knowledge that only *certain topics* are likely to be falsely reported. For example, news about politics (e.g., Brexit, U.S. election), leaders (e.g., Donald Trump, Pope Francis, Hillary Clinton), immigration (e.g., the "Bowling Green massacre")[2] or race (e.g., police shootings of Black men; the Black Lives Matter movement) might raise a flag and cause people to be cautious. Conversely, information that is less sensational (e.g., new scientific discoveries that do not have mass appeal) is more likely to be accurate. Thus, the content of news reports can act like a specific warning; the veracity of some items of information needs to be questioned whereas other information can be accepted at face value.

Naturally, a similar type of post-warning specificity could occur in forensic settings as well. For example, suppose a person witnesses a fight between two people, Greg and Joe. After the altercation, Joe's friends, who were present at the time, gather around the eyewitness and argue that Joe was not to blame – it was all Greg's fault – but introduce

---

[1] Warned participants may adopt certain strategies to limit the effect of misinformation in the absence of event memory such as avoiding familiar items on a memory test for fear that they are familiar for the wrong reason. However, as the original event memory has been destroyed under the overwriting hypothesis (i.e., only the misleading detail resides in memory), these strategies will be of limited value in moderating the effect of misinformation across different tests.

[2] The Bowling Green massacre was supposedly a terrorist attack referred to by U.S. Counselor to the President Kellyanne Conway in interviews with various media sources in early 2017. Reference to the massacre was intended to justify President Trump's proposed travel and immigration ban that affected travellers from several predominantly Muslim countries. However, the massacre never occurred.

misinformation in the process. When in court later on, the eyewitness may be vigilant about answering questions specifically about Joe's involvement in the altercation because he is aware that Joe's friends may have influenced him with misinformation.[3] Conversely, the eyewitness knows that he didn't receive any misinformation about Greg, so extra vigilance answering questions about him is not necessary. Compare this scenario to a second one in which an eyewitness receives a general warning prior to giving testimony to be careful about answering questions accurately on the stand. In a sense, the eyewitness in the first scenario has been post-warned about questions specifically to do with Joe's involvement in the altercation, but not about other types of questions, whereas the warning in the second scenario is more general. The question we address in the current research is, compared to a general warning, how effective are specific warnings at reducing the effect of misinformation on later memory performance.

Processes Involved in General and Item-Specific Post-Warnings

How do post-warnings generally affect remembering, and how can specific post-warnings amplify the beneficial effects on memory performance? Eyewitness testimony involves the conversion (Tulving, 1983) of pertinent memory information into a statement, such as an answer in a memory test. In principle, conversion includes a very broad range of processes, but of particular interest here are memory search as well as monitoring and control processes. Eyewitnesses would have to generate candidate answers and monitor their likelihood of being accurate (Koriat & Goldsmith, 1996), which potentially also involves monitoring the sources of candidate answers (Lindsay & Johnson, 1989). Control options include the volunteering or withholding of answers (Koriat & Goldsmith, 1996), testifying or not testifying answers (Higham et al., 2011), regulating the grain size or plurality of answers

[3] Alternatively, the judge may even caution the eyewitness about answering questions specifically about Joe because s/he is aware that Joe's friends spoke to the eyewitness afterwards.

(Goldsmith, Koriat, & Pansky, 2005; Luna, Higham, & Martín-Luengo, 2011; Luna & Martín-Luengo, 2012) or – in forced-choice recognition tests – choosing between provided responses.

Crucially, in eyewitness misinformation designs, these processes will differ substantially as a function of (1) the presence or absence of post-warnings and (2) the specificity of the post-warning. This is because post-warnings shape people's task representation, that is, their understanding of the memory task at hand and the necessary strategies to perform well (Blank, 1998; see also Lane, Roussel, Villa, & Morita, 2007, for a related approach). *Without* a post-warning, the task is (deceptively) simple: Drawing on a default consistency assumption (Blank, 1998), people will search memory for just one detail relevant to a test question, and will accept any familiar detail as the answer – which of course may be the misleading detail. Source monitoring is minimal, as the two sources of information (the original event and the post-event account) are assumed to be consistent; that is, it would be sufficient to place the remembered detail within the general situational context. Any post-warning about misinformation will potentially undermine this consistency assumption and as a consequence create a different task representation. Endorsing the most familiar item or the first item that comes to mind may be no longer sufficient; instead, people will need to search for potentially two contradictory details as candidate answers. Moreover, source monitoring becomes critical, as the sources of the details now have implications for their likely accuracy.

In short, post-warnings have the potential to change the task representation from a simple but problematic one (i.e., search for *one* familiar detail and report it as the answer: search-and-accept), to a complex but enlightened one (i.e., search for two contradictory details and monitor their sources to decide about their likely accuracy: search-and-discriminate). As the latter would help to weed out some inaccurate answers, performance is

expected to improve with a post-warning. Existing research supports this analysis. For instance, Echterhoff et al. (2005) found increased source monitoring after a post-warning; warned participants took longer to make memory decisions and rated their event memories higher on memory characteristics (e.g., visual details, vividness, and clarity of spatial context; Johnson, Foley, Suengas, & Raye, 1988). Further, compared to standard recognition tests, source-monitoring tests have been shown to reduce or even eliminate misinformation effects (e.g., Lindsay & Johnson, 1989).

General post-warnings, as argued above, induce a search for *potentially* two different details. What happens, however, if this search is unsuccessful and only one item comes to mind (which might be the misleading detail)? In such a case, any subsequent monitoring likely depends on what the witness/participant knows or believes regarding the overall presence of misinformation in the situation. If they assume the amount of misinformation to be small relative to the number of questions asked in the memory test, they may take the absence of memory for two contradictory details as evidence that there was no misinformation (similar to the impact of subjective theories about the memorability of items; Strack & Bless, 1994). As a consequence, they may forego any further source monitoring to validate the accuracy of the remembered detail. That is, general post-warnings may produce only lax monitoring overall, potentially allowing some misinformation to "slip through the net."

Consider item-specific post-warnings now, which indicate exactly those questions on the memory test for which misinformation had been presented earlier. Such post-warnings will optimize the task representation such that witnesses/participants will adopt a search-and-discriminate approach for misleading items and can safely adopt a search-and-accept approach for control items. This, first and foremost, means stricter monitoring and control, but it should also have consequences for memory search. When only one detail comes up

initially for a designated two-detail question, people may search memory harder for a second detail, and in some cases this may be successful. In short, item-specific post-warnings should help to reject misleading details and (perhaps to a lesser degree) retrieve original event details, hence improving memory performance across the board.

Finally, note that our analysis of the effects of general and item-specific post-warnings pointed to parallel expected effects on memory performance and metacognition (i.e., both improved memory performance and a reduced signature pattern of overconfidence in misinformation). This is because these effects are mediated through respective (less or more effective) task representations, which – as we argued above – have consequences for both memory and metamemory processes. We tested these ideas in two experiments.

Experimental Overview

We present two experiments that investigated the moderating influence of general and item-specific post-warnings on the misinformation's effect on memory performance and confidence. In Experiment 1, two groups both received a general post-warning. The *general-warning group* received just this post-warning but the *specific-warning group* received the general post-warning as well as item-specific post-warnings that informed participants whether each test question was a misleading question or a control question with a clear explication of what this meant. Experiment 1 also probed how perceived discrepancies between details impacted on memory performance. While Experiment 1 used a standard two-alternative forced-choice (2AFC) recognition test to assess memory performance, Experiment 2 employed a cued-recall test to rule out an alternative interpretation of the Experiment 1 specific-warning group results.

Experiment 1

In this experiment, participants studied a series of slides depicting a murder scene. Following each slide, participants read a brief narrative about that slide that contained one

piece of misinformation. Later, a 2AFC recognition memory test assessed participants' memory for the event details. Crucially, we investigated the effect of two types of post-warnings. Half of the participants were given just a general post-warning immediately prior to the memory test about the presence of misinformation in the narratives. The other half received the general post-warning as well as an item-specific post-warning. To implement the latter post-warning type, test questions were color-coded according to whether each question queried misleading (red) vs control (green) items.

Experiment 1 also explored the role of discrepancy detection in combination with a general post-warning. Several studies have noted that discrepancy detection plays an important role in moderating the misinformation effect because it puts participants in a position to discount misleading details (e.g., Blank, 1998; Higham 1998; Pohl, Schumacher, & Friedrich, 1993; Schooler & Loftus, 1986; Tousignant, Hall, & Loftus, 1986). Our interest here, however, was more in determining the outcome if participants failed to detect any discrepancy but still received a post-warning. As argued above, lack of subjective evidence of the presence of two discrepant details may give participants license to adopt an inadequate search-and-accept task representation, rendering a general post-warning ineffective in such cases. By contrast, the item-specific post-warning should still ensure an adequate search-and-discriminate task representation even in cases where no discrepancy was detected, leading to better memory performance than in the general-warning group.

Method

Participants. A total of 48 students from the University of Southampton participated individually in the experiment in exchange for course credits. Ages ranged from 18 to 56 years (M = 21.65 years, SD = 8.53 years). Twenty-four participants (23 females, 1 male) were assigned to the general-warning group and 24 (23 females, 1 male) to the specific-warning group.

Design and Materials. The "crime scene" consisted of 15 slides (digital photographs) and 15 corresponding narratives showing/describing a staged murder. The slides showed the perpetrator's car leaving the crime scene, the victim's home, a knife, and the victim's body. The narratives contained 30 critical details, two pertaining to each slide in the crime-scene sequence. One version of each critical detail (misinformation) misrepresented the detail in the slide (e.g., *bungalow* was mentioned in the narrative when in fact a *two-storey house* had been shown in the slide). The other version (control) either omitted the misinformation or described the detail in neutral form (e.g., *building*). The presence/absence of misinformation was varied within-subjects: For any given participant, half the critical details (one per slide) occurred in their misleading form and half in their control form, with the assignment of critical details to the control vs misleading forms counterbalanced across participants. The photographs that made up the slide sequence remained the same and were presented in the same order in the two counterbalanced formats.

A 30-item two-alternative recognition memory test was constructed and made into booklets. Each booklet contained a page of instructions followed by five pages containing a total of 30 questions. For each question, there were spaces to write an answer (A or B), a confidence rating about the accuracy of each response (50% [guess] - 100% [very confident]), and a decision about testifying (Y/N).[4] Each question on the test queried one critical detail and the two choices for each question were the correct event detail (e.g., *two-storey house*) and the misleading detail (e.g., *bungalow*). The questions appeared in chronological order, starting with questions about slide 1 and ending with questions about slide 15. Across questions, option A vs B represented the correct answer 14 vs 16 times,

[4] Following Higham et al. (2011), a side issue explored in both of our experiments was the impact of a testify option, as a supplementary and ecologically valid index of confidence, on memory performance and confidence. The findings largely paralleled the confidence findings and are therefore not reported.

respectively.

Procedure. Participants entered the lab and were seated in front of an Apple 24-inch iMac computer, which was used to present the crime-scene slides and narratives using Apple Keynote software. The first two slides contained instructions that informed participants that they would be shown a series of slides and written descriptions depicting a murder scene. They were informed that they should study the slides and descriptions closely. They were also told that it was vital that any information they provide in the experiment be accurate and informative and to ask any questions before proceeding. Participants then viewed the slides one at a time. Eight seconds were allotted to study each slide, after which the screen went blank, and then a written narrative appeared which described the details of the slide. The narratives corresponding to each slide ranged in length depending on the amount of detail depicted. Longer display times were implemented for longer narratives so that all participants could finish reading them (range: 15-25 seconds).

After the slide show and narratives had been presented, participants were given a Sudoku puzzle to complete for 10 minutes as a distractor task. They were then administered a test booklet containing a set of instructions on the first page (which included a post-warning – see below) and the questions for the recognition memory test on subsequent pages. While answering questions on the test, participants were asked to imagine that they were in a courtroom. They were instructed to answer all questions. A 50-100% confidence rating was also required.[5] After finishing the test, all participants were asked, "Did you notice any discrepancies between the pictures presented and the slide narratives?" If participants indicated "yes" to this question, they were asked to "…go back over the questionnaire and

[5] As participants manually entered their confidence ratings on a blank space, they were free to ignore instructions and provide confidence ratings lower than 50%. This happened infrequently, though; 3% of the ratings in Experiment 1 (6% in the general-warning group and 0% in the specific-warning group) were in this range. (Experiment 2 used a 0-100% confidence rating instruction.)

put a tick in the right hand margin corresponding to the questions that you recognized as having such a discrepancy."[6]

Post-warnings. Participants in both the general-warning and specific-warning groups were post-warned about the presence of misinformation in the narratives. In particular, the following instructions were printed on the first page of the memory test booklet, which participants read immediately prior to completing the memory test:

> *IMPORTANT: The narratives that you read earlier contained some inaccurate details. Don't assume that if you can remember a detail from the narrative, that it is guaranteed to be correct. To perform well on the test, you need to accurately remember what happened on the slides, which may or may not correspond to the account in the narratives.*

In addition to this general post-warning, participants in the specific-warning group were informed on an item-by-item basis which 15 test questions pertained to details for which there was misinformation presented earlier (misleading questions) and which 15 questions did not (control questions). This was achieved by printing the former question type in red typeface and the latter question type in green typeface and informing participants about the association between color and question type in the instructions given just prior to the memory test. Specifically, the following directions were printed on the first page of the memory-test booklet that contained the rest of the instructions:

---

[6] Discrepancies may have been detected during narrative encoding and/or later when the two discrepant details were explicitly presented to participants on the 2AFC recognition test. In contrast to our methodology that required participants to identify discrepancies at test, research on noticing and recollecting change, discussed in more detail later in this article, has typically required participants to identify discrepancies as they are first presented (e.g., Jacoby, Wahlheim, & Kelley, 2015). Requiring participants to identify discrepancies as soon as they are presented rather than later during testing means that no detected discrepancies are forgotten. However, it was not possible to follow this procedure in our design because alerting participants to the presence of discrepancies prior to narrative encoding would have constituted a pre-warning rather than a post-warning. Thus, although we admit that our procedure may have missed some discrepancy detection (e.g., some discrepancies detected during narrative encoding may have been forgotten by the time the test was written), as will become apparent, discrepancy-detection decisions made during the test were still very informative about the underlying processes.

PLEASE NOTE: There are 30 questions in total, each with two alternative answers. Fifteen of these questions relate to details about which you have been misinformed. In other words, a narrative that you read contained misleading information about that detail, so you have to be very careful when answering these questions. For these questions, one alternative is correct (i.e., it appeared only in the slides) whereas the other is incorrect (i.e., it was read about only in the narrative). The other 15 questions relate to details about which you have received no misinformation. In other words, the narrative did not contain misleading information about that detail. For these questions, one alternative is correct (i.e., it appeared only in the slides) whereas the other is incorrect (i.e., it is a new detail). To help you answer the questions correctly and make decisions about which answers to use in your testimony, misinformation questions are written in RED, whereas non-misleading questions are written in GREEN.

Test questions all appeared in black (consistent with the rest of the questionnaire) for participants in the general-warning group.

Results and Discussion

Some analyses required excluding a few participants because of empty or undefined cells. For example, if accuracy for a particular participant was 100%, then it was not possible to compute mean confidence for incorrect answers. Because many of our analyses were repeated-measures, the number of participants contributing data to different analyses varied (e.g., fewer participants were likely to contribute data to all cells in larger analyses involving many conditions compared to smaller analyses involving only a few conditions). The number of missing cases for each analysis is indicated in footnotes throughout the Results and Discussion sections. For the tables, the means and standard errors are based on all available data for each cell and may vary slightly from the means yielded from analyses for which

participants were excluded.

We first conducted analyses on all responses followed by further analyses that separated items according to whether a discrepancy was or was not detected.

Memory performance. Mean accuracy is shown in Table 1. A 2 (question type: misleading, control) × 2 (group: specific-warning, general-warning) mixed Analysis of Variance (ANOVA) on accuracy revealed a significant effect of question type, $F(1,46) = 15.42$, $MSE = 0.02$, $p < .001$, $\eta_p^2 = .25$ which was qualified by a question type by group interaction, $F(1,46) = 6.60$, $MSE = 0.02$, $p = .013$, $\eta_p^2 = .13$. Accuracy was higher for control items ($M = .76$, $SEM = .02$) than misleading items ($M = .65$, $SEM = .03$), but as shown in Table 1, this difference only existed in the general-warning group, $F(1,23) = 25.31$, $MSE = 0.02$, $p < .001$, $\eta_p^2 = .52$. The misinformation effect was eliminated when specific post-warnings were made available, $F < 1$. The main effect of group was not significant, $F(1,46) = 2.29$, $MSE = 0.03$, $p = .14$, $\eta_p^2 = .05$.

Confidence. Mean confidence is shown in Table 2. The data were initially analyzed with a 2 (question type: misleading, control) × 2 (group: specific-warning, general-warning) × 2 (accuracy: correct, incorrect) mixed ANOVA, with group as the only between-subjects factor.[7] It revealed main effects of question type, $F(1,42) = 5.07$, $MSE = 76.41$, $p = .03$, $\eta_p^2 = .11$, and accuracy, $F(1,42) = 101.98$, $MSE = 133.97$, $p < .001$, $\eta_p^2 = .71$, as well as a two-way interaction between question type and group, $F(1,42) = 9.15$, $MSE = 76.41$, $p = .004$, $\eta_p^2 = .18$. There was also a marginal two-way interaction between question type and accuracy, $F(1,42) = 4.05$, $MSE = 65.49$, $p = .05$, $\eta_p^2 = .09$. However, all these effects were qualified by a significant three-way interaction between question type, accuracy, and group, $F(1,42) = 9.07$, $MSE = 65.49$, $p = .004$, $\eta_p^2 = .18$.

---

[7] Four participants (one in the general-warning group and three in the specific-warning group) were dropped from this analysis because they made no errors in one of the cells.

To interpret the three-way interaction, we ran separate 2 (question type: misleading, control) × 2 (accuracy: correct, incorrect) repeated-measures ANOVAs on mean confidence in each group. The general-warning group ANOVA revealed main effects of question type, $F(1,22) = 13.19$, $MSE = 84.53$, $p = .001$, $\eta_p^2 = .38$, and accuracy, $F(1,22) = 44.05$, $MSE = 183.77$, $p < .001$, $\eta_p^2 = .67$. Confidence was higher for misleading items ($M = 74$, $SEM = 2$) than control items ($M = 67$, $SEM = 3$) and it was higher for correct responses ($M = 80$, $SEM = 2$) than incorrect responses ($M = 61$, $SEM = 3$). More interesting, there was a significant interaction between question type and accuracy, $F(1,22) = 13.64$, $MSE = 63.46$, $p = .001$, $\eta_p^2 = .38$. The interaction occurred because misinformation boosted confidence in incorrect responses, $F(1,22) = 16.30$, $MSE = 121.04$, $p = .001$, $\eta_p^2 = .43$, whereas it had no effect on correct responses, $F < 1$. Thus, the signature pattern of misinformation was observed in the general-warning group.

By contrast, there was no such interaction in the second 2 × 2 repeated-measures ANOVA on mean confidence in the specific-warning group, $F < 1$, only a main effect of accuracy, $F(1,20) = 72.35$, $MSE = 79.18$, $p < 001$, $\eta_p^2 = .78$. As with the previous analysis, correct responses ($M = 83$, $SEM = 1$) were assigned higher confidence than incorrect responses ($M = 67$, $SEM = 2$). This result, coupled with the fact that no misinformation effect was obtained on accuracy in the specific-warning group, indicates specific warnings eliminated the signature pattern of misinformation.[8]

---

[8] One could argue that specific warnings did not reduce confidence in wrong responses to misleading items (general-warning: $M = 65$, $SEM = 3$; specific-warning: $M = 67$, $SEM = 3$). Instead, it increased confidence in wrong control judgments (general-warning: $M = 54$, $SEM = 4$; specific-warning: $M = 68$, $SEM = 3$). However, in our view, some caution should be exerted in interpreting absolute confidence means in this way (rather than relative patterns) because it fails to take into account response bias (see Higham, Zawadzka, & Hanczakowski, 2016, for detailed discussion). It is likely that the specific warning caused confidence assignments to become more relaxed compared to the general warning, which would have increased both the control and misleading confidence means. If this difference in response bias is coupled with a genuine decrease in subjective confidence for wrong responses to misleading items in the specific-warning group, the result could be confidence means for misleading items that are approximately equal between the groups and control means that are different (just as we observed: see Table 2).

Discrepancy detection and post-warning effectiveness. Recall that at the end of the memory test, participants were required to mark any test questions that corresponded to a noticed discrepancy between the narratives and the slides. We used these data to further explore the effectiveness of general and item-specific post-warnings. As argued in the introduction, general post-warnings are ambiguous in that they leave it to the participants to decide which questions on the test the post-warning applies to and the task representation that would be adequate. If people failed to detect a discrepancy for a given test question, they should be more inclined to just search for one detail and accept it, possibly falsely (if it is a misleading detail). By contrast, detecting a discrepancy between the detail in the event and the one in the narrative should trigger a search-and-discriminate task representation and make people more resistant to misinformation by, for example, invoking more careful source monitoring. Thus, the task representation that is adopted for a particular test question in the general-warning group may depend to a large extent on discrepancy detection.

This logic does not apply to item-specific post-warnings, however, because, by their very nature, specific post-warnings already provide adequate task representations for both misleading and control questions, such that participants need not rely on the presence or absence of discrepancy detection to (mis-)specify them. In short, particularly the *absence* of discrepancy detection for misleading items should carry the risk of task misspecification and subsequent performance and monitoring deficits in the presence of general but not item-specific post-warnings. It is worth noting, though, that accuracy could still be poor for misleading items if a discrepancy was missed, even in the specific-warning group. For example, if the misleading detail was the only one recognized and this detail was misattributed to the event, then errors would result. However, we do not anticipate the effect of discrepancy detection in the specific-warning group to be as large as that observed in the general-warning group.

Hence, our analysis strategy was to focus on memory performance and monitoring for these *no discrepancy detected* (NDD) cases and to contrast them with cases where a discrepancy had been detected (*discrepancy detected* or DD cases; see Table 1). Overall, the vast majority of participants responded affirmatively to the Y/N question about whether any discrepancies were detected (general warning: 92%; specific warning: 88%). At the item level, the incidence of correct discrepancy detection (i.e., for misleading questions) was 38% in both the general- and specific-warning groups (both *SEMs* = 4). The false discrepancy detection rate (i.e., for control questions) was too low to statistically analyze (10% and 1% in the general- and specific-warning groups, respectively). Therefore, our subsequent re-analyses focused exclusively on misleading questions.[9]

In the general-warning group, responses to NDD misleading questions were substantially less accurate than responses to DD misleading questions (Table 1). Indeed, mean NDD accuracy was below chance, +95% confidence limit = .49, indicating that participants not only had their accuracy impaired by misinformation if a discrepancy was not detected, but they *preferred* the misleading detail over the event detail. By comparison, DD accuracy was very high – even higher than control accuracy, a point to which we return below. The same general pattern of better DD than NDD accuracy was present in the specific-warning group as well, even though participants knew the appropriate task representation. We attribute this residual difference in accuracy to source-monitoring failures. However, the drop in NDD accuracy compared to DD accuracy in the specific-warning group was not as great as in the general-warning group. Indeed, NDD accuracy differed between the groups, $F(1,46) = 9.02$, $MSE = 0.05$, $p = .004$, $\eta_p^2 = .16$. In short, failing to detect a discrepancy made people vulnerable to misinformation even in the presence of a specific

[9] All participants were included in these analyses. For those participants who indicated on the overall Y/N question that they failed to detect any discrepancies, all their test questions were coded as NDD.

post-warning, but the vulnerability was much worse if only a general post-warning was provided.

In contrast to the poor performance for misleading NDD items, accuracy for misleading DD items was near ceiling (Table 1) and *higher* than control accuracy. This observation was confirmed statistically: accuracy for misleading DD items exceeded control accuracy in both the general and specific warning groups, $F(1,21) = 7.51$, $MSE = 0.02$, $p = .012$, $\eta_p^2 = .26$ and $F(1,20) = 31.10$, $MSE = 0.01$, $p < .001$, $\eta_p^2 = .61$, respectively.[10] This finding is potentially interesting given recent research demonstrating that if participants notice and recollect change in classical retroactive (and proactive) interference paradigms, facilitation may be observed instead of interference (e.g., Jacoby et al., 2015; Jacoby, Wahlheim, & Yonelinas, 2013; Putnam, Wahlheim, & Jacoby, 2014; Wahlheim, 2014, 2015).

However, before elaborating further on either this facilitation for misleading DD items or the impairment for misleading NDD items, we considered it necessary to eliminate the possibility of item-selection artifacts. For example, it is quite plausible that NDD vs DD items are ones for which event memory is poor vs good, respectively, and it is this variation in event memory that is the reason for the accuracy difference between the item types, not the variation in the rate of discrepancy detection per se. Indeed, a correlational analysis showed that control performance – as an uncompromised (by misinformation) measure of memory strength for original details – was correlated with discrepancy detection across the 30 test items; $r = .45$ and $r = .49$ in the general- and specific-warning groups, respectively (both significantly above zero, $p < .05$).

To investigate this possibility, we followed others in the change-recollection literature (e.g., Jacoby et al., 2015; Putnam, Sungkhasettee, & Roediger, 2017; Putnam et al., 2014) and

[10] Five participants (two vs three in the general- vs specific-warning groups, respectively) were excluded from these analyses because they indicated that they had detected no discrepancies.

conducted a hierarchical regression analysis at the level of items. In this model, accuracy for misleading items was the dependent variable, whereas the predictors were (a) accuracy for control items (as a measure of item memory), (b) the difference in the discrepancy detection rate between misleading and control items (the difference taken to control for guessing), and (c) the interaction between these two variables. Control item accuracy was entered first, followed by the discrepancy detection rate, and then the interaction. If the difference in accuracy for DD vs NDD items was entirely due to differential event memory, then the discrepancy detection variable would not account for any additional variance once the control-item accuracy was entered on the first step. However, if discrepancy detection per se had an effect on performance above and beyond variations in event memory, then discrepancy detection would account for some additional unique variance. Because the data patterns were similar between the groups (i.e., NDD < control < DD, with similar rates of correct discrepancy detection), we pooled them to increase power.

The regression analysis indicated that the total amount of variance explained by the three predictors was $R^2 = .63$. As expected, control-item accuracy entered on step 1 was a significant predictor of misleading-item accuracy, $\Delta R^2 = .52$, $p < .001$. More critically, discrepancy detection entered on step 2 also accounted for a significant amount of additional unique variance, $\Delta R^2 = .11$, $p = .009$. Finally, the amount of variance accounted for by the interaction between these variable entered on step 3 was not significant, $\Delta R^2 = .00$, $p = .90$. Thus, the regression analysis indicates that although item selection played a role in producing facilitation for DD items and impairment for NDD items, it by no means accounted for the full effect; detecting discrepancies also had a unique effect on performance.

A critic might argue that the association between discrepancy detection and memory performance for misleading items in this analysis may not be due to discrepancy detection causing better event memory, but rather it reflects the reverse causal relationship. On this

view, there may be fluctuations in attention that vary on a participant-by-participant basis. For example, a random attentional lapse for one participant could interrupt his/her ability to encoding the details of a slide even though that slide resulted in excellent item memory for most other participants. Conversely, another participant might idiosyncratically focus on an item that is missed by most other people. Under most circumstances, such attentional fluctuations would simply be considered statistical noise. However, in the present context, idiosyncratic fluctuations may be problematic in that they may independently affect discrepancy detection and later performance on the memory test for that participant. Ultimately, the critic argues, there is only one causal variable, item memory, which takes two forms in our regression analysis: a stable, item-based component which is captured by average control accuracy, and an idiosyncratic one which is captured by the discrepancy-detection variable. Critically, by this account, discrepancy detection per se has no causal influence on item memory or performance on the memory test, a conclusion that is completely at odds with our interpretation of the regression results.

Although our data do not permit us to eliminate this account absolutely, we do not believe that random attentional fluctuations occurred often enough to fully account for the added effect of discrepancy detection in our regression analysis. First, control accuracy was highly correlated across items between the two warning conditions, $r = .74$, $p < .001$. As noted above, idiosyncratic attentional fluctuations would introduce statistical noise into the estimates of control accuracy. If these fluctuations occurred with any regularity, statistical noise would be high, resulting in a low correlation between these variables. Instead, the fact that this correlation was high, despite the different procedures implemented between the groups, suggests that any attentional fluctuations were few and far between. Second, as we discuss in more detail below, a growing body of research across different paradigms, including the misinformation paradigm (e.g., Putnam et al., 2016), has indicated that covert

retrieval of original memories during discrepancy detection can have a facilitative effect (Jacoby et al., 2013, 2015; Putnam et al., 2014; Wahlheim, 2014, 2015). Hence, we believe it would be imprudent to attribute the added effect of discrepancy detection in our hierarchical regression analysis entirely to random attentional fluctuations. Nonetheless, future research investigating the causal role of discrepancy detection in the misinformation paradigm might implement different procedures to more firmly establish causality (e.g., experimentally manipulate the likelihood of discrepancy detection for the same set of items).

Our final analysis was to investigate the relationship between discrepancy detection and confidence.[11] Inspection of Table 2 reveals a complementary picture to accuracy in terms of confidence for correct and incorrect NDD answers. In the general-warning group, the signature misinformation pattern was preserved for incorrect NDD answers; that is, there was higher confidence assigned to incorrect answers to misleading NDD questions compared to control questions, $F(1,22) = 18.02$, $MSE = 127.67$, $p < .001$, $\eta_p^2 = .45$. By contrast, the *absence* of this pattern was replicated (with respect to the analysis on all responses) in the specific-warning group, $F(1,20) = 1.03$, $MSE = 118.53$, $p = .32$, $\eta_p^2 = .05$. Further, in the general-warning group, there was a complete breakdown of discrimination between correct and incorrect misleading NDD answers, $F < 1$, but not in the specific-warning group, $F(1,21) = 33.56$, $MSE = 30.76$, $p < .001$, $\eta_p^2 = .62$. That is, the item-specific post-warning not only improved memory performance but also memory monitoring, compared to the general post-warning.

Summary and interpretation. In the general-warning group, we found a misinformation effect on accuracy, accompanied by a boost to confidence for incorrect responses – the signature pattern of misinformation. Thus, similar to several other reports

[11] Between one and three participants were dropped in each analysis because of empty cells. DD answers were not included in this analysis because there were too few incorrect DD answers to make this meaningful.

(e.g., Belli et al., 1994; Greene et al., 1982; Higham, 1998; Higham et al., 2011; Zaragoza & Lane, 1994), an influence of misinformation was still evident despite participants' general awareness of its presence. In contrast, performance in the specific-warning group was much better – the item-specific post-warning completely eliminated both the misinformation effect on accuracy and the effect on confidence. Essentially, once participants knew which questions were which, there was no discernable effect of misinformation at all.

However, these beneficial effects of specific post-warnings were less pronounced if participants failed to detect a discrepancy between the detail in the event and the detail in the narrative. Under those circumstances, a robust misinformation effect was observed, even after controlling for item-selection artifacts, an effect we attribute to problems monitoring the source of misleading details that were retrieved without a corresponding event detail. Although these source-monitoring problems likely occurred in the general-warning group as well, they were exacerbated by an inappropriate task representation. Participants provided only with a general post-warning and who recognized only one detail in response to a test question likely came to believe that they were answering a control question and continued search efforts were unnecessary. As a result, they endorsed the misleading details frequently and with high confidence.

Finally, an unexpected finding was that retrieval of event details was *facilitated* by misinformation if a discrepancy was detected. Again, this effect persisted even after controlling for item-selection artifacts. Retroactive facilitation has recently been shown to occur in the classical retroactive interference paradigm (as well as the proactive interference paradigm; e.g., Jacoby et al., 2015) and Putnam et al. (2017) have recently demonstrated retroactive facilitation in a misinformation paradigm. We believe this finding is important at both a theoretical and applied level and so we return to it again in the General Discussion.

Experiment 2

The purpose of Experiment 2 was to address two issues related to the use of a 2AFC recognition test in Experiment 1. First, there is a possible alternative interpretation of the observed efficacy of item-specific post-warnings: As a simple shortcut for generating answers to test questions, participants could have decided, for some of the misleading questions, to just switch their answers from their initially preferred response to the other one. That is, upon learning (through the specific post-warning) that there *might* be a problem with the detail they remembered, they simply opted for the other alternative in some cases. If what they initially remembered was the misleading detail, this would have resulted in an apparent but not genuine improvement of memory accuracy in the specific-warning group.

Second, it has long been known that recognition, by virtue of being supported by the most efficient retrieval cue – the item itself – is not as vulnerable to retroactive interference (from post-event misinformation, for instance) as recall (e.g., Postman & Stark, 1969). Hence, even if there was a genuine improvement in *recognition* accuracy, item-specific post-warnings may prove less efficient in less supported (in terms of retrieval cues), but perhaps more ecologically valid, retrieval situations. To address these issues, Experiment 2 used a cued-recall procedure.

Method

Participants. A total of 44 students from the University of Southampton participated individually in the experiment in exchange for course credits. Ages ranged from 18 to 33 years (M = 22.84 years, SD = 3.49 years). Twenty-two participants (15 females, 7 males) were assigned to the general-warning group and 22 (11 females, 11 males) to the specific-warning group.

Design, materials and procedure. The design, materials and procedure in Experiment 2 were mostly the same as in Experiment 1 except that (1) participants received a cued-recall test after the post-warning, (2) confidence ratings were made on a 0-100% rather

than a 50-100% scale, and (3) no discrepancy-detection decisions were gathered at the end of the experiment.[12] For the cued-recall task, instead of choosing between two response alternatives for each test question, a space was provided for participants to write their answer. Confidence ratings and testify decisions were collected as before. Participants were explicitly instructed to guess if they did not know the answer to a question; this instruction was used to avoid losing too many responses for the confidence and monitoring analyses.

The general post-warning was the same as in Experiment 1. However, to accommodate the cued-recall task, the specific post-warning had to be amended slightly as follows:

> *PLEASE NOTE: There are 30 questions in total. Fifteen of these questions relate to details about which you have been misinformed. In other words, a narrative that you read contained misleading information about that detail, so you have to be very careful when answering these questions. The other 15 questions relate to details about which you have received no misinformation. In other words, the narrative did not contain misleading information about that detail. To help you answer the questions correctly and make decisions about which answers to use in your testimony, misinformation questions are written in RED, whereas non-misleading questions are written in GREEN.*

Coding of recall answers. Cued-recall responses were coded into five categories: (1) critical-event detail (corresponding to the correct response alternative in the 2AFC test used

---

[12] Discrepancy-detection decisions were not implemented in this experiment because we used cued-recall testing rather than 2AFC recognition as in Experiment 1. For 2AFC recognition, it is clear which details were to be judged for discrepancies because they were presented to participants as recognition alternatives. For example, one test question was "In photograph 1, what was at the end of the road?" and participants chose between "two-storey building" (event detail) and "bungalow" (misleading detail). However, neither of these responses was necessarily made on the cued-recall test. For example, a legitimate response would have been "house" (counted as noncritical-correct detail; see section on coding). Because the details to be assessed for discrepancies were not well specified in cued recall, the discrepancy detection data would have been difficult or even impossible to interpret.

in Experiment 1), (2) noncritical-event detail (i.e., an event detail that was technically correct, but which was not specifically the critical-event detail), (3) critical-misleading detail (corresponding to the misleading 2AFC response alternative), (4) noncritical-incorrect detail (any incorrect detail other than the critical-misleading detail), and (5) unclassifiable response. For example, in response to the cue "In photograph 1, what was at the end of the road?" the responses "two-storey building," "house," bungalow," "a cat," and "dunno" would constitute categories 1-5, respectively. After pooling the data from the general- and specific-warning groups, categories 1-5 accounted for 41%, 19%, 14%, 21% and 5% of all answers provided, respectively. Our analyses below focus primarily on categories 1 and 3.

Results and Discussion

Memory performance. Table 3 shows the mean proportion of control and misleading questions with critical-event details and critical-misleading details as responses in the cued-recall task. The former details counted as one type of correct response whereas the latter counted as one type of error. A 2 (question type: control, misleading) × 2 (group: specific-warning, general-warning) mixed ANOVA on the proportion of critical-event details recalled yielded no significant main effects, largest $F(1,42) = 2.37$, $MSE = 0.01$ $\eta_p^2 = .05$, but there was a significant interaction, $F(1,42) = 4.51$, $MSE = 0.01$, $p = .040$, $\eta_p^2 = .10$. Follow-up tests on the interaction revealed little difference in the proportions of correctly recalled critical-event details for control and misleading questions in the general-warning group, $F < 1$, whereas in the specific-warning group, the proportion was *greater* for misleading questions than control questions, $F(1,21) = 6.05$, $MSE = 0.01$, $p = .023$, $\eta_p^2 = .22$ (Table 3).

The analogous 2 × 2 ANOVA on the proportion of critical-misleading details falsely recalled found significant main effects of question type, $F(1,42) = 29.36$, $MSE = 0.01$, $p < .001$, $\eta_p^2 = .41$, and group, $F(1,42) = 10.15$, $MSE = 0.01$, $p = .003$, $\eta_p^2 = .20$. False recall was higher for misleading questions ($M = .19$, $SEM = .02$) than control questions ($M = .09$, $SEM =$

.01) and it was higher in the general-warning group (M = .17, SEM = .01) than in the specific-warning group (M = .11, SEM = .01). However, both these main effect were qualified by a significant interaction, $F(1,42) = 11.40$, $p = .002$, $MSE = 1.84$, $\eta_p^2 = .21$. Follow-up tests on the interaction indicated a large misinformation effect in the general-warning group, $F(1,21) = 37.68$, $MSE = 0.01$, $p < .001$, $\eta_p^2 = .64$, but no comparable effect in the specific-warning group, $F(1,21) = 2.14$, $MSE = 0.01$, $p = .158$, $\eta_p^2 = .09$ (Table 3).

Confidence. Participants' mean confidence in correctly recalled critical-event details and falsely recalled critical-misleading details in the general- and specific-warning groups is shown in Table 4. As in Experiment 1, mean confidence was analyzed with a 2 (question type: control vs. misleading) × 2 (response: correct vs. incorrect) × 2 (group: general-warning, specific-warning) mixed ANOVA with group as the only between-subjects factor.[13] It revealed only a main effect of response, $F(1,30) = 55.81$, $MSE = 437.01$, $p < .001$, $\eta_p^2 = .65$. Unsurprisingly, correct responses (M = 86, SEM = 2) were assigned higher confidence than incorrect responses (M = 59, SEM = 4). No other main effects or interactions were significant, all Fs < 1. The signature pattern of misinformation – greater confidence in incorrect responses for misleading as opposed to control questions – was still descriptively present in the general-warning group, but it did not reach significance, $F < 1$. Also, confidence in incorrect responses to misleading questions was descriptively higher in the general-warning group than in the specific-warning group, but again not significantly so, $F(1,39) = 2.00$, $MSE = 708.48$, $p = .165$, $\eta_p^2 = .05$.[14]

Additional analyses. We conducted Experiment 2 primarily to eliminate the possibility that a response-switching strategy was the cause of specific warnings having such

[13] Twelve participants (five vs seven in the general- vs specific-warning groups respectively) were dropped from this analysis due to empty cells.
[14] For these last two analyses, ten participants were dropped from the first (five from each group) and three participants were dropped from the second (all in the specific-warning group) because of empty cells.

a profound effect on memory performance in Experiment 1. Compared to the 2AFC recognition task used in Experiment 1, for which two candidate answers were explicitly presented for every question, no candidate answers were explicitly presented in the cued-recall task used in Experiment 2. Consequently, it was not as straightforward for participants to switch away from the more familiar (narrative) detail to a less familiar (slide) detail when specifically warned about the presence of misinformation and improving memory accuracy as a result.

However, the critic could argue that, although it is not as straightforward, response switching could still potentially occur in cued recall as well. Participants may, for example, covertly retrieve both the more familiar misleading detail along with the less familiar event detail in response to the question. If participants are specifically warned that a particular question is dangerous but they are unsure about the source of each candidate response, they may strategically elect to report the less familiar event detail, which would lead to better recall performance. This criticism is important to reject because it potentially could explain both the enhanced recall of event details, and the lower rate of falsely recalling misleading details, for misleading questions compared to control questions in the specific-warning group.[15]

To address this criticism, we conducted two analyses. The first was an item analysis for which we correlated two variables in the specific-warning group. The first variable was the amount of recall facilitation for critical-event details that each question yielded in its misleading form compared to its control form (i.e., misleading recall proportion minus control recall proportion for each question). The second variable was the number of different

[15] Although the goal of strategic response switching in cued recall is similar to that in the 2AFC task, there is an important difference. In 2AFC, the event detail may not be retrieved and have no familiarity *at all*; however, participants may still select it simply to avoid the familiar (narrative) detail. In contrast, the event detail in cued recall must be retrieved for participants to be able to switch to it. In other words, response switching requires retrieval of the event detail in some form to operate in cued recall, whereas it does not in 2AFC.

candidate responses that were produced for each question across participants (set size). We reasoned that any given participant would be more likely to be entertaining several candidate responses for questions with large vs small set sizes. Furthermore, response switching was likely to produce greater facilitation for questions with small set sizes rather than large ones because there would be fewer competitors to interfere with reporting the event detail once the misleading detail was discounted. In other words, response switching predicts a negative relationship between these variables (greater set size, less facilitation). However, contrary to this prediction, the results of this analysis revealed a *positive* correlation between the variables, $r = .45$, $p = .013$.

Our second analysis focused on recall of the noncritical-event details in the specific-warning group. As noted above, participants sometimes produced details on the recall test that were technically correct because they were shown in the slides, but they were not *critical*-event details (e.g., recalling "house" instead of the critical-event detail "two-storey building"). If response switching was the cause of excellent performance in the specific-warning group, then recall of noncritical-event details for misleading questions should be augmented relative to control questions, just as it was for the critical-event details. However, this was not the case; recall of noncritical-event details to misleading questions ($M = .15$, $SEM = .02$) was *impaired* relative to control questions ($M = .25$, $SEM = .02$), $F(1,21) = 9.49$, $MSE = 0.012$, $p = .006$, $\eta_p^2 = .31$. Thus, specific warnings did not just facilitate reporting of *any* correct information – the enhancement was specific to *critical*-event information. Coupled with the results of the previous analysis, this analysis allowed us to safely eliminate response switching as the basis of our results.

Summary and interpretation. Similar to Experiment 1, we found higher levels of misinformation endorsement with general as opposed to item-specific post-warnings in Experiment 2. Indeed, specific post-warnings completely eliminated the misinformation

effect even though the retrieval cues were less efficient (i.e., cued recall instead of 2AFC recognition). Moreover, strategic response switching was not the cause of this excellent memory performance in the specific-warning group. First, response switching was made more difficult by using a cued-recall task in Experiment 2. Second, subsequent analyses eliminated the possibility that participants overcame this difficulty by strategically reporting less familiar covertly-generated candidate responses.

An additional unanticipated effect of specific post-warnings in Experiment 2 was that correct recall of critical-event details was *greater* for misleading questions than control questions, a pattern that did not occur with a general post-warning (Table 3). We return to this surprising finding in the General Discussion. Finally, Experiment 2 replicated the beneficial effect of item-specific post-warnings on confidence: the signature misinformation pattern was eliminated, whereas it was still at least descriptively present in the general-warning group.

General Discussion

The purpose of this research was to investigate the impact of post-warning specificity on memory performance and monitoring in the eyewitness misinformation paradigm. We conducted two experiments, the first using a standard 2AFC recognition procedure and the second using cued recall. Similar to several previous studies (e.g., Belli et al., 1994; Frost et al., 2002; Greene et al., 1982; Higham, 1998; Higham et al., 2011), the general post-warning administered in Experiment 1 was not very effective at reducing the effect of misinformation on either accuracy or confidence. Instead, the signature pattern of misinformation observed in other research (e.g., Loftus et al., 1989; Luna & Migueles, 2009) was preserved in the general-warning group: reduced memory accuracy and inappropriately high confidence when misinformation was erroneously accepted.

By contrast, the item-specific post-warning completely eliminated both the

misinformation effect on accuracy and the exaggerated confidence in endorsed misinformation. Experiment 2 further established that the effect of item-specific post-warnings is not limited to peculiarities of 2AFC recognition procedures – which could invite simple heuristics such as switching responses for dangerous questions – but extends to a cued-recall setting where such heuristics are of less use. In the remainder of this discussion, we will address a number of particularly noteworthy findings before drawing some general conclusions.

Discrepancy Detection and Misleading Details

Why was the general post-warning administered in Experiment 1 not very effective at reducing the misinformation effect? That is, what differentiates our study from some other studies that did find full elimination of the misinformation effect using general post-warnings (e.g., Christiaansen & Ochalek, 1983; Highhouse & Bottrill, 1995; Oeberst & Blank, 2012)? It is clear from our experiments that failure to detect discrepancies was at the heart of the problem in the general-warning group in Experiment 1; accuracy for misleading NDD questions, for which no discrepancy was detected, was half that of control questions and significantly below chance. This large misinformation effect for misleading NDD items was preserved even after controlling for item-selection artifacts. These data suggest that the signature pattern of misinformation found in the complete data set described above was primarily driven by extremely poor performance (coupled with inappropriately high confidence) on misleading questions for which discrepancy detection failed.

Performance for misleading details in both the general- and specific-warning groups is depicted in Figure 1. (Ignore the information associated with “E”– the event detail – for the moment.) Given the importance of discrepancy detection, Figure 1 distinguishes between cases where discrepancy detection was indicated at test and cases where it was not. Figure 1 shows that if a discrepancy was successfully detected (left-hand side of Figure 1), there was

an appropriate task representation which led to low endorsement of misleading details and low confidence assigned to the few misleading details that were endorsed (Outcome A). This outcome was the same regardless of the type of post-warning and corresponds to the outcome for DD items in both warning groups of Experiment 1 and analogous items in both warning groups of Experiment 2.[16]

In contrast, if a discrepancy detection was not indicated at test (right-hand side of Figure 1), the task representation and the ultimate outcome depended on the post-warning type. We suspect in a lot of these cases the misleading detail was the only one retrieved, but it was retrieved lacking source information. A general post-warning was not effective enough for participants to adopt an appropriate task representation and to be cautious about endorsing this single detail. As a result, it was fully endorsed with high confidence (Outcome C). This outcome corresponds to the results for NDD items in the general-warning group of Experiment 1 and analogous items in the same group in Experiment 2.

On the other hand, if there was no indication of discrepancy detection at test and participants were specifically post-warned, participants adopted an appropriate task representation and they only endorsed the misleading detail with moderate frequency and assigned moderate confidence to it (Outcome B). Endorsement and confidence was tempered under these circumstances because, although there may have been a candidate response for the question (the misleading detail in many cases), participants were aware that *two* discrepant details were associated with the question, even though they could not explicitly identify them. As a result of this more adequate task representation, a continued search may have ensued which on some occasions may have been successful, leading to somewhat higher

[16] Although the cued-recall test in Experiment 2 did not allow us to explicitly identify DD and NDD items (see Footnote 11), the procedure in Experiment 2 was the same as Experiment 1 up to the point of testing. Consequently, we assume that discrepancy detection occurred during narrative encoding in Experiment 2 at approximately the same rate as Experiment 1.

memory accuracy compared to NDD items in the general-warning group. However, if the search was unsuccessful, participants may have guessed, which would moderate both accuracy and confidence. This outcome corresponds to the NDD items in the specific-warning group of Experiment 1 and analogous items in the same group in Experiment 2.

More principally, the argument would be that (post-)warnings are only effective to the degree that participants' subjective impression of the potential to make memory errors (implied by the post-warning) matches the actual potential to make those errors. Post-warning *specificity* (in our case implemented as item-specific post-warnings) contributes to post-warning effectiveness by narrowing this subjective-objective gap and informing participants' task representations and ensuing retrieval strategies (e.g., failing to recollect a discrepancy after being specifically warned called for continued memory search). We think that exploring different aspects of such (mis)matches between subjective and objective memory task contexts could be a worthwhile avenue for future research.

Discrepancy Detection and Event Details

Although failing to detect discrepancies for misleading items had a disastrous effect on accuracy in Experiment 1, particularly if participants were only provided with a general warning, substantial benefits were observed if discrepancies *were* detected. Accuracy on misleading DD items in Experiment 1 was near ceiling and exceeded control accuracy by a substantial degree (Table 1). In all likelihood, this accuracy advantage was not solely due to the fact that memory for event details was good for DD items. Although the control accuracy exerted a significant effect in the hierarchical regression analysis, suggesting that item selection played a partial role in this facilitative effect, the analysis also pointed to an additional unique contribution of discrepancy detection.

Why would presenting misleading information to participants be associated with such excellent performance? As we noted above, Jacoby, Wahlheim and colleagues (e.g., Jacoby

et al., 2013, 2015; Putnam et al., 2014; Wahlheim, 2014, 2015) have investigated analogous facilitation effects in classical interference paradigms. In the retroactive interference version of this paradigm, which is closest to the paradigm used to study misinformation effects, participants first study word pairs in an initial list and then study a second list in which the stimulus in the pair is presented with a different response (i.e., A-B, A-D). Memory for the initial pairing is then tested (A-?, for which the correct response is "B") and the typical finding is that memory is impaired compared to a control condition (A-B, C-D). However, Jacoby et al.'s (2015) interesting novel finding was that if participants noticed that the response paired with the stimulus had changed between the first and second list (or to use our lingo, they detected a discrepancy), and they successfully recollected that change at test, recall performance in the interference condition exceeded that in the control condition. For example, in their Experiment 1, recall accuracy in the A-B, C-D control condition was 40%. However, if the experimental context was conducive to detecting and recollecting change, recall accuracy in the A-B, A-D interference condition was significantly higher at 50%.

Jacoby et al. (2015) interpreted such facilitative effects within a *recursive-remindings* framework (e.g., Hintzman, 2011). A central tenet of this framework is that noticing change (i.e., detecting discrepancies) requires that the original, pre-changed stimulus (or stimulus pair) be covertly retrieved. Hence, the process of detecting change engenders retrieval practice (or a spaced covert repetition) of the original stimulus, which is well-known to enhance memory (e.g., Carrier & Pashler, 1992; Roediger & Karpicke, 2006), even if the retrieval is covert (e.g., M. A. Smith, Roediger, & Karpicke, 2013). Applying this logic to the misinformation paradigm, which is a special case of the retroactive interference paradigm, the message is that discrepancy detection does not just serve to limit endorsements of the misleading detail, but it can also enhance memory for the original event. This enhancement is important because it suggests that, like the classical retroactive interference paradigm with

word pairs, there may actually be two effects produced by exposure to misinformation: interference if a discrepancy is not detected but facilitation if it is.

Although the importance of discrepancy detection in limiting vulnerability to misinformation has been documented in the past (e.g., Blank, 1998; Higham, 1998; Schooler & Loftus, 1986; Tousignant et al., 1986), very little research has focused on the facilitative effect on event memory that misinformation can have when it is coupled with discrepancy detection during narrative encoding. One exception is Oeberst & Blank (2012) who found a memory advantage for event details in the misleading condition in cases where the task representation was very clearly specified. They attributed this effect partly to discrepancy detection during narrative encoding leading to deeper processing of the event detail (see also Blank, 2005, for related effects in a classical interference paradigm). More recently, Putnam et al. (2017) conducted two experiments on the misinformation effect using a three-alternative recognition test consisting of the event detail, the misleading detail, and a new detail. Similar to our Experiment 1 results, they found that detecting change led to greater endorsement of the event detail, and lower endorsement of the misleading detail, compared to control items. These studies, together with our current results, suggest that facilitation of event memory due to misinformation may be fairly common but potentially masked in many studies by interference effects (i.e., the net effect of misinformation on performance is typically negative). However, if performance is made conditional on discrepancy-detection decisions (as in Putnam et al. and our Experiment 1) or if an appropriate task representation is greatly emphasized (as in Blank, 2005, Oeberst & Blank, 2012, and in the specific-warning group of the current Experiment 2), then facilitation will be observed.

Facilitation due to covert retrieval practice is depicted on the left-hand side of Figure 1 as “E: enhanced memory due to covert retrieval practice during narrative encoding.” It is associated with Outcome A, which requires discrepancy detection. Outcome A corresponds

to the misleading DD questions in both post-warning groups of Experiment 1 (and to analogous but not explicitly identified misleading questions in the specific-warning group of Experiment 2). In all these cases, retrieval practice of event details led to better-than-control performance for the misleading items.

Another potential facilitative effect of misinformation may have been at work in Experiment 2. In that experiment, we observed that cued recall of event details in the specific-warning group – but not in the general-warning group – was greater for misleading items than control items (see Table 3). As the two warning groups were treated identically prior to the memory test, the rate of discrepancy detection during narrative encoding (and associated Outcome A; Figure 1) was likely comparable between the two warning groups and therefore cannot explain facilitation in one group but not the other.

It is worth noting at the outset that this enhanced event memory for misleading vs control items is completely at odds with the "overwriting" or "destructive updating" hypothesis (e.g., Loftus, 1979a, 1979b; Loftus & Loftus, 1980; Loftus et al., 1978). By this hypothesis, the original event memory would have been overwritten by the misinformation, so there was no way that facilitation due to the receipt of misinformation could have occurred instead. Furthermore, even if there was a way of explaining the facilitation, the destructive updating hypothesis still leaves unexplained why the facilitation occurred in the specific-, but not the general-warning group.

The question at this juncture is: If it was not discrepancy detection or destructive updating that caused the facilitation, then what caused it? We see an important mechanism contributing to this facilitative effect on event detail recall in Experiment 2 as extended memory search at test specifically for misleading questions in the specific-warning group. In the specific-warning group, the adequate task representation conveyed by the specific warning motivated participants to continue searching memory for two details if they failed to

detect a discrepancy for misleading questions (right-hand side of Figure 1). In contrast, there was no need to extend the search if only one detail came to mind for control questions because participants were informed that there is only one associated detail. This enhanced searching specifically for misleading questions may have sometimes been successful but sometimes may have caused retrieval of the misleading detail. However, as we have argued above, we suspect the appropriate task representation in the specific-warning group altered not just the time and effort devoted to searching memory to answer misleading questions, but also enhanced source monitoring (i.e., the knowledge that there were two discrepant details invoked more stringent source-monitoring processes). The net result was better event-detail recall for misleading questions compared to control questions in the specific-warning group of Experiment 2 (i.e., Outcome B in Figure 1). In the general-warning group, by comparison, participants likely had a dysfunctional search-and-accept task representation for misleading NDD items that undermined the motivation to continue searching memory if a discrepancy was not detected and only one detail was retrieved. The net result was lower event-detail recall for misleading vs control questions (Outcome C in Figure 1), because the one retrieved detail would often have been the misleading detail (e.g., due to recency) and the – potentially available – event detail was never retrieved, as the search was not continued.

Generally, then, the group difference in event detail recall for misleading items reflects the differential prevalence of Outcomes B and C. The total level of facilitation (i.e., overall misleading minus control performance) observed in the groups (+8% and -2% in the specific- and general-warning groups, respectively) reflects a combination of Outcomes A and (mostly) B in the specific-warning group and a combination of Outcomes A and (mostly) C in the general-warning group (with the contribution of A being constant because of the identical procedure up to the point of testing). Hence, the overall misleading facilitation effect observed in the specific-warning group of Experiment 2 may reflect both encoding-

based and retrieval-based facilitation, whereas the absence of overall facilitation in the general-warning group likely reflects a mixture of encoding-based facilitation and retrieval-based interference effects (i.e., misleading detail endorsement) for different items that approximately cancelled each other out.

Beyond demonstrating facilitation effects in the misinformation paradigm, the present findings also add to the nascent memory facilitation literature in another respect. Unlike the effects found with traditional interference designs and with Putnam et al.'s (2017) misinformation paradigm, our facilitation effects occurred *without* any explicit instructions to detect discrepancies prior to narrative encoding (see e.g. Jacoby et al., 2015, for typical instructions). Rather, because of the typical consistency assumption in misinformation studies (Blank, 1998), participants likely did not expect any change at all (they were only alerted to the possibility of change in the post-warning about misinformation). Therefore, any change/discrepancy detection occurred spontaneously. It would be interesting to determine in future research if spontaneous change detection is more or less facilitative than guided change detection. We suspect it may be the former, as spontaneous detection is likely more surprising and therefore should lead to more elaboration of the changed elements.

Conclusion

Our research adds to a growing body of research that demonstrates the effectiveness of (some) post-warnings against misinformation (Blank & Launay, 2014). It extends previous research by focusing on post-warning specificity and examining the processes underlying general and item-specific post-warnings. By contrasting these two types of post-warnings, we discovered a potential Achilles heel of general post-warnings – their ambiguity in terms of the adequate task representation and retrieval strategy for individual test items. General post-warnings are – paradoxically – not necessarily general, in that they do not convey an adequate representation and effective strategy by default, across the board. Rather, their

effects seem to materialize locally, when supported by memory for original details and previous discrepancy detections. In the absence of such support, general post-warnings may be completely ineffective. Thus, the varied success of general post-warnings could be partially explained by differences in these associated features and processes.

The effectiveness of specific post-warnings, in contrast, suggests that due caution when answering questions about specific topics or people can potentially overcome the negative effects of misinformation. This finding comes as some relief given the prevalence of misinformation in the form of "fake news" in today's "post-truth" society. As we noted in the Introduction, specific post-warnings might take the form of questioning the veracity of memories pertaining to particular topics or people that may be associated with misinformation. An interesting avenue for future research would be to investigate whether self-generated specific warnings that are topic- or person-based are as effective as externally generated ones such as those used in our current research.

In a more general perspective, the differential post-warning effects featured in this article illustrate the complexity of the interaction between task instructions, stored memory information, misinformation, and metacognitive processes. Not too long ago, misinformation was believed to have simple, straightforward effects on memory for witnessed event details (e.g., memory impairment: Loftus, 1991; response biases: McCloskey & Zaragoza, 1985). The new picture that has emerged more recently highlights, in addition to such influences, a variety of processes that intervene between memory retrieval and memory report (see our discussion of conversion and metacognitive monitoring and control processes in the Introduction to this article). In this new picture, external influences (such as misinformation) rarely have a direct, unmediated influence on memory. Rather, they are absorbed, along with other relevant information, in a constructive act of remembering. As a result, memory performance in the face of misinformation will be sometimes impaired (the typical case),

sometimes unaffected, and sometimes, when supported by discrepancy detection, even improved. Exploring the intricacies of the interplay between memory, testing conditions and task representations along the lines sketched in the present research will help understand this variability in outcomes.

References

Bekerian, D. A., & Bowers, J. M. (1983). Eyewitness testimony: Were we misled? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 9,* 139–145. doi:10.1037/0278-7393.9.1.139

Belli, R. F., Lindsay, D. S., Gales, M. S., & McCarthy, T. T. (1994). Memory impairment and source misattribution in postevent misinformation experiments with short retention intervals. *Memory & Cognition, 22*, 40-54. doi:10.3758/BF03202760

Blank, H. (1998). Memory states and memory tasks: An integrative framework for eyewitness memory and suggestibility. *Memory, 6,* 481-529. doi:10.1080/741943086

Blank, H. (2005). Another look at retroactive and proactive interference: A quantitative analysis of conversion processes. *Memory, 13,* 200-224. doi:10.1080/09608210344000698

Blank, H. & Launay, C. (2014). How to protect eyewitness memory against the misinformation effect: A meta-analysis of post-warning studies. *Journal of Applied Research in Memory and Cognition, 3,* 77-88. doi:10.1016/j.jarmac.2014.03.005

Carrier, M. & Pashler, H. (1992). The influence of retrieval on retention. *Memory and Cognition*, 20, 632-642. doi:10.3758/BF03202713

Chan, J. C. K., & LaPaglia, J. A. (2013). Impairing existing declarative memory in humans by disrupting reconsolidation. *Proceedings of the National Academy of Sciences*, *110*, 9309–9313. doi:10.1073/pnas.1218472110

Christiaansen, R. E., & Ochalek, K. (1983). Editing misleading information from memory: Evidence for the coexistence of original and postevent information. *Memory & Cognition, 11*, 467-475. doi:10.3758/BF03196983

Dodd, D. H., & Bradshaw, J. M. (1980). Leading questions and memory: Pragmatic constraints. *Journal of Verbal Learning and Verbal Behavior, 19*, 695-704.

doi:10.1016/S0022-5371(80)90379-5

Eakin, D. K., Schreiber, T. A., & Sergent-Marshall, A. (2003). Misinformation effects in eyewitness memory: The presence of memory impairment as a function of post-warning and misinformation accessibility. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*, 813-825. doi:10.1037/0278-7393.29.5.813

Echterhoff, G., Hirst, W. & Hussy, W. (2005). How eyewitnesses resist misinformation: Social postwarnings and the monitoring of memory characteristics. *Memory & Cognition, 33*, 770-782. doi:10.3758/BF03193073

Frost, P., Ingraham, M., & Wilson, B. (2002). Why misinformation is more likely to be recognised over time: A source monitoring account. *Memory, 10*, 179-185. doi:10.1080/09658210143000317

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

Greene, E., Flynn, M. S., & Loftus, E. F. (1982). Inducing resistance to misleading information. *Journal of Verbal Learning and Verbal Behavior, 21*, 207-219. doi:10.1016/S0022-5371(82)90571-0

Higham, P. A. (1998). Believing details known to have been suggested. *British Journal of Psychology, 89*, 265-283. doi:10.1111/j.2044-8295.1998.tb02684.x

Higham, P.A., Luna, K., & Bloomfield, J. (2011). Trace-strength and source-monitoring accounts of accuracy and metacognitive resolution in the misinformation paradigm. *Applied Cognitive Psychology*, *25*, 324-335. doi:10.1002/acp.1694

Higham, P. A., Zawadzka, K., & Hanczakowski, M. (2016). Internal mapping and its impact on measures of absolute and relative metacognitive accuracy. In J. Dunlosky & S. K. Tauber (Eds.), *The Oxford Handbook of Metamemory* (pp. 39–61). New York, NY:

Oxford University Press. doi:10.1093/oxfordhb/9780199336746.013.15

Highhouse, S., & Bottrill, K. V. (1995). The influence of social (mis)information on memory for behavior in an employment interview. *Organizational Behavior and Human Decision Processes, 62*, 220-229. doi:10.1006/obhd.1995.1045

Hintzman, D. L. (2011). Research strategy in the study of memory: Fads, fallacies, and the search for the "coordinates of truth." *Perspectives on Psychological Science, 6*, 253–271. doi:10.1177/1745691611406924

Jacoby, L. L., Wahlheim, C. N., & Kelley, C. M. (2015). Memory consequences of looking back to notice change: Retroactive and proactive facilitation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*, 1282-1297. doi:10.1037/xlm0000123

Jacoby, L. L., Wahlheim, C. N., & Yonelinas, A. P. (2013). The role of detection and recollection of change in list discrimination. *Memory and Cognition, 41*, 638–649. doi:10.3758/s13421-013-0313-x

Johnson, M. K., Foley, M. A., Suengas, A. G., & Raye, C. L. (1988). Phenomenal characteristics of memories for perceived and imagined autobiographical events. *Journal of Experimental Psychology: General, 117*, 371-376. doi:10.1037//0096-3445.117.4.371

Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory. *Psychological Review, 103*, 490-517. doi:10.1037//0033-295X.103.3.490

Lane, S. M., Roussel, C. C., Villa, D., & Morita, S. K. (2007). Features and feedback: Enhancing metamnemonic knowledge at retrieval reduces source-monitoring errors. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33,* 1131–1142.

Lindsay, D. S. (1990). Misleading suggestions can impair eyewitnesses' ability to remember event details. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*, 1077-1083. doi:10.1037//0278-7393.16.6.1077

Lindsay, D. S., & Johnson, M. K. (1989). The eyewitness suggestibility effect and memory for source. *Memory & Cognition, 17*, 349-358. doi:10.3758/BF03198473

Loftus, E. F. (1979a). *Eyewitness testimony*. Cambridge, MA: Harvard University Press.

Loftus, E. F. (1979b). The malleability of human memory. *American Scientist, 67,* 312–320.

Loftus, E. F. (1991). Made in memory: Distortions in recollection after misleading information. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (pp. 187-212). San Diego: Academic Press. doi:10.1016/S0079-7421(08)60124-3

Loftus, E. F. (2005). Planting misinformation in the human mind: A 30-year investigation of the malleability of memory. *Learning & Memory, 12*, 361-366. doi:10.1101/lm.94705

Loftus, E. F., Donders, K., Hoffman, H. G., & Schooler, J. W. (1989). Creating new memories that are quickly accessed and confidently held. *Memory & Cognition, 17*, 607-616. doi:10.3758/BF03197083

Loftus, E. F., & Loftus, G. R. (1980). On the permanence of stored information in the brain. *American Psychologist, 35,* 409–420. doi: 10.1037/0003-066X.35.5.409

Loftus, E. F., Miller, D. G., & Burns, H. J. (1978). Semantic integration of verbal information into a visual memory. *Journal of Experimental Psychology: Human Learning and Memory, 4*, 19-31. doi:10.1037//0278-7393.4.1.19

Luna, K., Higham, P.A., & Martin-Luengo, B. (2011). The regulation of memory accuracy with multiple answers: The plurality option. *Journal of Experimental Psychology: Applied, 17,* 148-158. doi:10.1037/a0023276

Luna, K. & Martín-Luengo, B. (2012). Improving the accuracy of eyewitnesses in the presence of misinformation with the plurality option. *Applied Cognitive Psychology, 26*, 687-693. doi:10.1002/acp.2845

Luna, K., & Migueles, M. (2009). Acceptance and confidence of central and peripheral

misinformation. *The Spanish Journal of Psychology, 12*, 405-413. doi:10.1017/S1138741600001797

McCloskey, M., & Zaragoza, M. S. (1985). Misleading postevent information and memory for events: Arguments and evidence against memory impairment hypotheses. *Journal of Experimental Psychology: General, 114*, 1-16. doi:10.1037//0096-3445.114.1.1

Oeberst, A., & Blank, H. (2012). Undoing suggestive influence on memory: The reversibility of the eyewitness misinformation effect. *Cognition, 125,* 141-159. doi:10.1016/j.cognition.2012.07.009

Pohl, R.F., Schumacher, S., & Friedrich, M. (1993). The eyewitness misinformation effect: Distorted recollections based on contradictory information. In G. Strube & K.F. Wender (Eds.), *The cognitive psychology of knowledge* (pp. 33-52). Amsterdam: Elsevier Science Publishers B.V.

Postman, L., & Stark, K. (1969). Role of response availability in transfer and interference. *Journal of Experimental Psychology, 79*, 168-177. doi:10.1037/h0026932

Putnam, A. L., Sungkhasettee, V., & Roediger, H. L. (2017). When misinformation improves memory: The effects of recollecting change. *Psychological Science*, *28*, 36-46. doi: 10.1177/0956797616672268

Putnam, A. L., Wahlheim, C. N., & Jacoby, L. L. (2014). Memory for flip-flopping: Detection and recollection of political contradictions. *Memory & Cognition, 42,* 1198–1210. doi:10.3758/ s13421-014-0419-9

Rindal, E. J., DeFranco, R. M., Rich, P. R., & Zaragoza, M. S. (2016). Does reactivating a witnessed memory increase its susceptibility to impairment by subsequent misinformation? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*, 1544–1558. doi:10.1037/xlm0000265

Roediger, H. L. & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests

improves long-term retention. *Psychological Science, 17*, 249-255. doi:10.1111/j.1467-9280.2006.01693.x

Schooler, J. W., & Loftus, E. F. (1986). Individual differences and experimentation: Complementary approaches to interrogative suggestibility. *Social Behavior*, 105-112.

Smith, M. A., Roediger, H. L., & Karpicke, J. D. (2013). Covert retrieval practice benefits retention as much as overt retrieval practice. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39,* 1712–1725. doi:10.1037/a0033569

Smith, V. L., & Ellsworth, P. C. (1987). The social psychology of eyewitness accuracy: Misleading questions and communicator expertise. *Journal of Applied Psychology, 72*, 294-300. doi:10.1037//0021-9010.72.2.294

Strack, F., & Bless, H. (1994). Memory for nonoccurrences: Metacognitive and presuppositional strategies. *Journal of Memory and Language, 33,* 203-217. doi:10.1006/jmla.1994.1010

Tousignant, J. P., Hall, D., & Loftus, E. F. (1986). Discrepancy detection and vulnerability to misleading postevent information. *Memory & Cognition, 14*, 329-338. doi:10.3758/BF03202511

Tulving, E. (1983). *Elements of episodic memory.* Clarendon Press: Oxford.

Wahlheim, C. N. (2014). Proactive effects of memory in young and older adults: The role of change recollection. *Memory & Cognition, 42,* 950–964. doi:10.3758/s13421-014-0411-4

Wahlheim, C. N. (2015). Testing can counteract proactive interference by integrating competing information. *Memory & Cognition, 43,* 27–38. doi:10.3758/s13421-014-0455-5

Wright, D. B. (1993). Misinformation and post-warnings in eyewitness testimony: A new testing procedure to differentiate explanations. *Memory, 1*, 153-166. doi:10.1080/09658219308258229

Zaragoza, M. S., & Lane, S. M. (1994). Source misattributions and the suggestibility of eyewitness memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*, 934-945. doi:10.1037//0278-7393.20.4.934

Zaragoza, M. S., & Mitchell, K. J. (1996). Repeated exposure to suggestion and the creation of false memories. *Psychological Science, 7*, 294-300. doi:10.1111/j.1467-9280.1996.tb00377.x

Table 1

Mean Accuracy in Experiment 1 as a Function of Question Type and Post-Warning Group. Standard Errors Are Shown in Parentheses.

	Question Type			
	Control	Misleading		
Group	Overall	Overall	NDD	DD
General-warning	.78 (.03)	.59 (.04)	.39 (.05)	.90 (.03)
Specific-warning	.75 (.02)	.71 (.04)	.58 (.04)	.95 (.02)

Note: The overall mean for misleading questions is based on items for which no discrepancy was detected (NDD) and items for which discrepancy was detected (DD).

Table 2

Mean Confidence (%) in Experiment 1 as a Function of Question Type, Accuracy, and Post-Warning Group. Standard Errors Are Shown in Parentheses.

	Question Type		
	Control	Misleading	
Group and Accuracy	Overall	Overall	NDD
General-warning			
Correct	79 (2)	80 (2)	69 (3)
Incorrect	54 (4)	67 (3)	68 (3)
Specific-warning			
Correct	83 (1)	83 (1)	74 (2)
Incorrect	68 (3)	65 (3)	64 (2)

Note: NDD = misleading items for which no discrepancy was detected. Due to occasional empty cells, the means and standard errors are based on *N*s ranging from 22 to 24.

Table 3

Mean Proportion of Control and Misleading Questions with Critical-Event Details (Correct) and Critical-Misleading Details (Incorrect) as Responses in Experiment 2. Standard Errors Are Shown in Parentheses.

	Question Type	
Detail Type and Group	Control	Misleading
Critical-event details (correct)		
General-warning	.40 (.03)	.38 (.03)
Specific-warning	.39 (.03)	.47 (.04)
Critical-misleading details (incorrect)		
General-warning	.08 (.01)	.25 (.02)
Specific-warning	.09 (.01)	.13 (.02)

Table 4

Mean Confidence (%) in Experiment 2 as a Function of Question Type, Post-Warning Group, and Accuracy. Standard Errors Are Shown in Parentheses.

	Question Type	
Accuracy and Group	Control	Misleading
General-warning		
Correct	89 (1)	85 (2)
Incorrect	58 (6)	67 (5)
Specific-warning		
Correct	87 (3)	87 (2)
Incorrect	61 (8)	55 (7)

Note. The correct and incorrect answers taken into account for this analysis were the recalled critical-event and critical-misleading details pertaining to a test question (not any non-critical correct or non-critical incorrect answers). Due to occasional empty cells, the means and standard errors are based on *N*s ranging from 17 to 22.

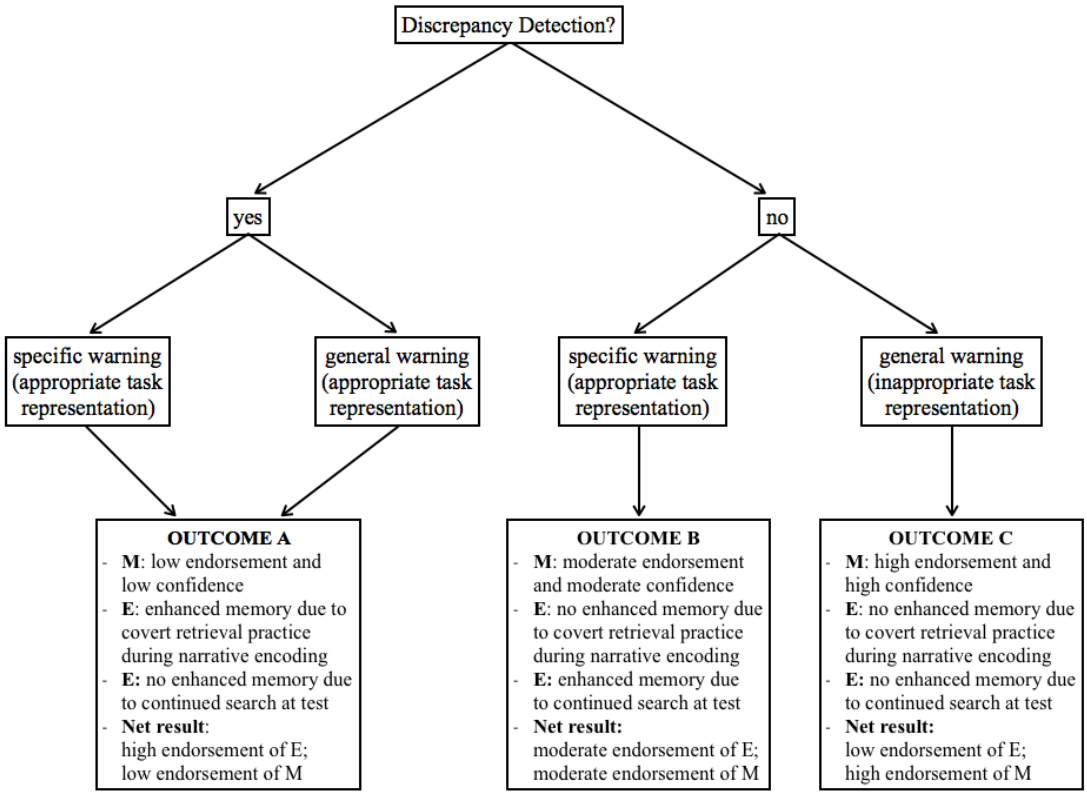

Figure 1. Flowchart of the underlying processes for misleading items leading to three potential outcomes in Experiments 1 and 2. In both experiments, discrepancy detection leads to fundamentally different results compared to no discrepancy detection. If a discrepancy is detected, regardless of whether participants are provided with a specific or general warning, misleading details (M) are associated with low endorsement and low confidence. Also, detecting discrepancies causes covert retrieval practice and enhanced memory of the event detail (E) (Outcome A). The net result is high endorsement of E and low endorsement of M. If no discrepancy is detected and there is a general post-warning, M is frequently endorsed with high confidence whereas there is no enhanced memory due to covert retrieval practice of E during narrative encoding (Outcome C). The net result is high endorsement of M and low endorsement of E. However, if discrepancy detection fails but participants are specifically warned, although there is no enhanced memory of E due to covert retrieval during narrative encoding, memory for E may be enhanced because participants conduct a more thorough search of memory at test. This thorough memory search may lead to greater retrieval of E relative to a general warning where the search may be aborted prior to retrieval of E (Outcome B). The net result is moderate endorsement of both E and M.