Retrieval-Enhanced Suggestibility: A Retrospective and a New Investigation

Jason C.K. Chan
Iowa State University, ckchan@iastate.edu

Krista D. Manley
Iowa State University, kmanley@iastate.edu

Kathryn Lang
Iowa State University

Follow this and additional works at: https://lib.dr.iastate.edu/psychology_pubs
Part of the Cognition and Perception Commons

The complete bibliographic information for this item can be found at https://lib.dr.iastate.edu/psychology_pubs/45. For information on how to cite this item, please visit http://lib.dr.iastate.edu/howtocite.html.

This Article is brought to you for free and open access by the Psychology at Iowa State University Digital Repository. It has been accepted for inclusion in Psychology Publications by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
Retrieval-Enhanced Suggestibility: A Retrospective and a New Investigation

Abstract
Eyewitnesses are often repeatedly interviewed about the same crime, and they might be exposed to misleading information between these interviews. Taking a memory test before being exposed to misinformation can ironically increase the likelihood that an eyewitness would fall prey to the misinformation effect—a finding termed Retrieval-Enhanced Suggestibility (RES). In this paper, we outline the motivation behind the original research on RES, summarize the theoretical explanations that have been used to explain this finding, and provide a review of the extant empirical findings. We then report the results of two new experiments that examined whether participants, upon being informed that they had provided inconsistent responses during the prior recall tests, could overcome the RES effect during a final recognition test in which the correct event detail and the misinformation were offered as response options. The results showed that RES persisted in this recognition test, and the level of inconsistencies shown by participants during the prior recall tests predicted their final recognition performance, such that more inconsistent witnesses were also more error-prone during the final recognition test—but this relation was stronger for misinformation-induced inconsistencies than for spontaneously-produced inconsistencies.

Keywords
Misinformation, Retrieval-enhanced suggestibility, Delay, Confidence, Eyewitness inconsistency, Recognition

Disciplines
Cognition and Perception | Psychology

Comments
Retrieval-Enhanced Suggestibility: A Retrospective and a New Investigation

Jason C.K. Chan
Krista D. Manley
Kathryn Lang*

*Iowa State University

Corresponding author:
Jason C.K. Chan

Department of Psychology
Iowa State University
Ames, IA 50011
USA

ckchan@iastate.edu
Abstract

Eyewitnesses are often repeatedly interviewed about the same crime, and they might be exposed to misleading information between these interviews. Taking a memory test before being exposed to misinformation can ironically increase the likelihood that an eyewitness would fall prey to the misinformation effect – a finding termed Retrieval-Enhanced Suggestibility (RES). In this paper, we outline the motivation behind the original research on RES, summarize the theoretical explanations that have been used to explain this finding, and provide a review of the extant empirical findings. We then report the results of two new experiments that examined whether participants, upon being informed that they had provided inconsistent responses during the prior recall tests, could overcome the RES effect during a final recognition test in which the correct event detail and the misinformation were offered as response options. The results showed that RES persisted in this recognition test, and the level of inconsistencies shown by participants during the prior recall tests predicted their final recognition performance, such that more inconsistent witnesses were also more error-prone during the final recognition test – but this relation was stronger for misinformation-induced inconsistencies than for spontaneously-produced inconsistencies.

*Keywords: misinformation; retrieval-enhanced suggestibility; delay; confidence; eyewitness inconsistency; recognition*
**General Audience Summary**

Eyewitness memory is malleable and can be altered by misleading suggestions, and this misinformation effect can be exacerbated by a prior interview. Here, we provide a brief theoretical and empirical review of the literature on this retrieval-enhanced suggestibility effect; we then examined whether these misinformation-induced errors would persist after participant witnesses were warned that they had provided inconsistent answers to the same question previously. Our results indicate that the misinformation effect persisted despite this warning. We also examined whether inconsistent witnesses are also “bad” witnesses – a prevalent belief held by lay people, law enforcement officials, and judicial actors. When inconsistent responses were produced spontaneously, the level of inconsistency was a weak predictor of eyewitness memory accuracy. However, when inconsistent responses were induced by misinformation, the level of inconsistency strongly predicted eyewitness accuracy. These results show that not all eyewitness inconsistencies are created equal, and the importance of preventing witnesses from being exposed to misinformation.

Word Count: 154
Retrieval-Enhanced Suggestibility: A Retrospective and a New Investigation

Repeated interviews are a fact of life in the legal proceedings of a criminal trial. Witnesses may be questioned by officers at the crime scene and/or at the police station, by attorneys during interviews and deposition, and finally, while producing their testimony in court. Recent research has demonstrated the potentially negative consequences of conducting multiple interviews on the accuracy of eyewitness memory reports. In particular, witnesses are sometimes more susceptible to misinformation if their memory for the witnessed event had been tested prior to misinformation exposure. This retrieval-enhanced suggestibility (RES) effect is typically revealed in a four-phase procedure. In Phase 1, participants view a witnessed event, which is usually a video or a slideshow that depicts a crime. In Phase 2, participants either receive a memory test for that original event (i.e., retrieval practice) or not (i.e., control). In Phase 3, participants are exposed to some pieces of misinformation, which are usually embedded in a narrative that purportedly recapitulates the witnessed event. In Phase 4, participants' memory for the original event is assessed again in a final test. The surprising result is that participants who have completed the initial test are sometimes more likely to incorporate misinformation into their final memory reports than participants who have not completed the initial test (Chan, Thomas, & Bulevich, 2009). The RES effect is notable because a substantial body of research has demonstrated that retrieval practice enhances, rather than impairs, subsequent memory performance (Roediger & Butler, 2011). Therefore, conditions that violate this assumption are of important theoretical interests.

In the following, we briefly describe the motivation behind the initial research on RES, the theoretical explanations that have been proposed to explain this counterintuitive result, and some of the more noteworthy empirical findings to date. We then present two experiments that investigate a new wrinkle in the research on eyewitness suggestibility in the context of RES –
i.e., whether the level of inconsistency in prior memory reports is indicative of subsequent eyewitness memory accuracy.

**Motivation Behind the Work that Led to the Discovery of RES**

In 2005, the first author (Chan) was conducting research that examined the beneficial effects of retrieval practice on later recall of both tested and nontested materials (Chan, McDermott, & Roediger, 2006), and Chan wondered if retrieval practice could be used as an intervention against the deleterious effects of misinformation for eyewitnesses. Chan had conducted a project on the misinformation effect as an undergraduate student (Lindsay, Allen, Chan, & Dahl, 2004), and felt that retrieval practice might be a “missing piece” in the traditional, three-phase misinformation design (which is similar to the four-phase design described above, but with the retrieval practice Phase 2 omitted). The logic was that upon viewing a crime, many eyewitnesses would recall details of the event soon after its occurrence. For example, one might briefly recall what happened to a 911 operator, provide a detailed statement to a police officer at the crime scene, or simply retell the event to a loved one. Due to the recency with which these recall “tests” occur after the witnessed event, they may help solidify one’s memory of the witnessed event before one is exposed to misleading information. Because having a stronger memory trace for the witnessed event should reduce one’s susceptibility to misinformation, the three-phase design, by virtue of omitting the initial “test,” may overestimate the influence of misinformation in many eyewitness situations (in which witnesses often recall the event prior to being exposed to misinformation). Based on this logic, Chan, together with Ayanna Thomas and John Bulevich, conducted the initial experiments that led to the discovery of RES – a result that was opposite to what the research team had predicted. Following the first experiment, Chan, Thomas, and Bulevich extended the results to older adults, and later completed an additional experiment that helped provide some much-needed theoretical analyses. This work ultimately culminated in a publication in *Psychological Science* (Chan et al., 2009).
Why Does Retrieval Exacerbate the Misinformation Effect?

A comprehensive examination of the theoretical mechanisms that underlie RES is beyond the scope of the present paper, but the “standard” explanation for RES includes two factors. The first, and most commonly supported explanation suggests that because retrieval practice can potentiate subsequent learning of new information (Chan, Davis, & Meissner, 2015; Pastotter & Bauml, 2014; Tulving & Watkins, 1974), prior testing may enhance participants’ ability to learn subsequently presented information – including misleading information. This enhanced learning of the misinformation is in turn manifested as heightened suggestibility (Chan et al., 2009; Chan & LaPaglia, 2011; Gordon & Thomas, 2014; Thomas, Bulevich, & Chan, 2010). We refer to this explanation as the test-potentiated learning (TPL) idea. Although the exact mechanisms by which retrieval strengthens subsequent learning are not yet clear, researchers have proposed several explanations, including optimization of encoding strategies, reduction of inattention/mind wandering, isolation of encoding contexts, etc. (Chan et al., 2015; Chan & McDermott, 2007; Chan, Davis, & Meissner, 2017; Pastotter & Bauml, 2014; Szpunar, Khan, & Schacter, 2013; Thomas et al., 2010).

The TPL explanation states that prior testing would strengthen encoding of the misinformation relative to control, thereby increasing its likelihood of being reported later. However, given that performing retrieval practice should also strengthen memories of the witnessed event, this benefit must be counteracted by the test-potentiated learning of the misinformation. Therefore, holding other factors constant (such as source monitoring and participants’ perceived accuracy of misinformation narrative), whether RES occurs or not should depend on the relative increase in strength between memories for the original event and the misinformation due to retrieval practice (see Figure 1 for an illustration). In general, this explanation suggests that because testing enhances learning of the misinformation, testing should increase recall probability of the misinformation relative to control. In tests that require participants to provide a single answer for a given question (e.g., cued recall), the heightened
response competition between the misinformation and the original event memories might also reduce recall probability of the original witnessed detail. However, in tests designed to minimize the influence of response competition such as modified-modified free recall (MMFR, in which participants can provide multiple answers to the same question) or free choice recognition, retrieval of the witnessed event detail should not be suppressed by the test-strengthened misinformation, because retrieval practice should enhance memories for both the witnessed event and the misinformation according to the TPL account.

Another explanation for RES suggests that recalling details of a witnessed event makes these memories more susceptible to interference from misinformation. Based on this notion, encountering misinformation shortly after participants have attempted to recall a relevant witnessed event detail would make subsequent retrieval of that witnessed event detail more difficult, relative to a condition in which one encounters misinformation without recent retrieval (Chan & LaPaglia, 2013). The logic of this account stems from the memory reconsolidation model, which asserts that reactivating a consolidated memory can, under some circumstances (Exton-McGuinness, Lee, & Reichelt, 2015; R. S. Fernandez, Boccia, & Pedreira, 2016; Krawczyk, Fernandez, Pedreira, & Boccia, 2017; Lopez et al., 2016), temporarily increase its malleability (Agren, 2014; Ecker, 2015), thereby rendering it more susceptible to modification. It is important to note that this account provides a supplementary, but not sufficient, explanation for RES. In particular, this explanation is best invoked when the timing parameters of an RES procedure conform to the consolidation-reactivation-reconsolidation disruption framework. Specifically, for misinformation to disrupt reconsolidation of the original event memory, that memory should first be consolidated before it is reactivated, and the misinformation should be presented during the hypothetical reconsolidation period. Studies that are designed to properly test this account thus require a two- or three-phase procedure, in which retrieval practice of the original event occurred at least one day after its encoding (such that the original memory could consolidate), and the interfering (mis)information is presented shortly after this retrieval practice
(for a review of human declarative memory studies that satisfy these parameters, see Scully, Napper, & Hupbach, 2016). Consequently, studies that are conducted in a single experimental session do not provide a critical test of this account (Chan et al., 2009; Rindal, DeFranco, Rich, & Zaragoza, 2016).

Lastly, another potential explanation for the RES effect is that retrieval enhances suggestibility because it increases misinformation acceptance (Belli, 1989), rather than producing actual changes to the underlying memory representations of the original event and the misinformation (Chan & LaPaglia, 2011; Rindal et al., 2016). The idea is that taking a prior memory test increases the likelihood that participants would treat the misinformation-embedded narrative as providing corrective feedback; that is, testing increases the perceived credibility of the misinformation. This account was most clearly articulated by Rindal and colleagues, although it has yet to be explicitly tested. Similar to the reconsolidation account, this notion may provide a supplementary, but insufficient, explanation for RES, given that memory tests that minimize the influence of misinformation acceptance (e.g., MMFR, source-free recognition test, narrative recall test, to be described in the empirical review to follow) still reveal RES. In addition, retrieval practice sometimes reduces, rather than increases, suggestibility (Gabbert, Hope, Fisher, & Jamieson, 2012; LaPaglia & Chan, 2012; Pansky & Tenenboim, 2011), a finding that should not occur if retrieval practice increases the likelihood for people to treat the misinformation narrative as corrective feedback.

It is important to note here that the above explanations are not mutually exclusive, and they may all, under different circumstances, contribute to the occurrence/magnitude of RES. Again, among processes described by these explanations, test-potentiated learning of the misinformation is likely sufficient to produce RES, whereas reconsolidation disruption and test-enhanced misinformation acceptance might contribute to, or augment, the RES effect in some situations.
Some Notable Empirical Findings of RES

Since its initial discovery, the RES effect has been demonstrated under a variety of experimental situations. In this section, we briefly review some of the major findings. For exposition purposes, we organize this literature review using the encoding, storage, and retrieval heuristic.

**Encoding variables.** There are typically at least two encoding events in studies that examine the misinformation effect: one when participants encounter the initial witnessed event and another when participants encounter the materials through which misinformation is delivered. A variety of factors have been investigated, including variations in the length and modality of the witnessed event, the type of misinformation, the method of delivering misinformation, and the participant population.

**Modality of Witnessed Event.** RES has been shown using a variety of materials, including video materials ranging from about 5 min (Wilford, Chan, & Tuhn, 2014) to 40 min in length (Chan et al., 2009). Researchers have also demonstrated RES using slideshows instead of videos (Rindal et al., 2016). In addition, two recent studies have shown RES-like effects using factual knowledge (Mullet, Umanath, & Marsh, 2014; Pashler, Kang, & Mozer, 2013), and another study reported an RES effect with word pairs (Weber, 2012). Despite these results, several studies have also revealed that retrieval practice can reduce the negative influence of misinformation (Gabbert et al., 2012; Huff, Davis, & Meade, 2013; LaPaglia & Chan, 2012; Pansky & Tenenboim, 2011). In Huff et al., the materials were six pictures (shown at a rate of 18 sec) depicting scenes with various objects in it, and the participants were given supplementary (i.e., objects that were absent from the original pictures), rather than contradictory, misinformation. In LaPaglia and Chan, a 45-sec video depicted a theft, and the critical item concerned the appearance of the perpetrator (with the misinformation being that the perpetrator had a goatee). In Gabbert et al., the witnessed event was a roughly 2-min video of a bank
robbery, and participants were introduced to three pieces of misinformation. It is not yet clear why these studies demonstrated a benefit of retrieval practice rather than RES, but a common feature across these studies is that the witnessed event was quite short, and so the misinformation was also delivered via a short narrative. As will be described below, LaPaglia (2013; see also LaPaglia & Chan, 2013) showed that the way in which misinformation is presented can have a profound impact on whether prior testing would increase or decrease suggestibility.

**Types of Misinformation.** Most studies presented misinformation that contradicted peripheral event details, and these studies typically showed a robust RES effect (e.g., Chan & LaPaglia, 2011; Gordon, Thomas, & Bulevich, 2015). However, Wilford et al. (2014) included misinformation for both central and peripheral details, and they found that RES was more likely to happen for peripheral details than for central details. The results regarding central details were somewhat ambiguous given the near ceiling accurate recall performance for these items. Nonetheless, these authors reported a small but significant RES effect for the central items when they combined their data across three experiments. More recently, Gordon and colleagues (2015) compared the RES effect for contradictory and supplementary misinformation, and they found RES for both, but the effect appeared to be larger for supplementary misinformation than for contradictory misinformation (see also Davis, Fernandez, & Chan, 2017).

**Methods of Delivering Misinformation.** Chan and LaPaglia (2013) presented misinformation either in the context of the witness event (i.e., a terror attack) or in the context of an unrelated event (i.e., a drug bust), and they found an RES effect in only the former condition. LaPaglia and Chan (2013) presented misinformation in a narrative or in individual questions. Somewhat surprisingly, retrieval increased suggestibility when the misinformation was presented in a narrative, but it reduced suggestibility when the misinformation was presented via questions. Similarly, Pansky and Tenenboim (2011) showed a beneficial effect of retrieval practice when misinformation was embedded in questions (instead of a narrative). In addition, Memon and
colleagues (2010b) found a testing effect for forcibly fabricated misinformation following an initial Cognitive Interview.

Why did a testing effect, instead of RES, occur in these aforementioned studies? One possibility, as alluded to previously, is related to the context under which misinformation is encoded. In a factorial design, LaPaglia (2013) investigated this issue by presenting misinformation in a narrative or in questions, but she also manipulated an additional variable – whether the misinformation was presented with or without contextual information. When misinformation was presented with contextual information, the sentences/questions that contained misinformation were embedded into a narrative that recapitulated the story of the witnessed event. When misinformation was presented without contextual information, the sentences/questions that contained misinformation were presented in an isolated manner, without being embedded in the story-like narrative of the witnessed event. Critically, LaPaglia found that it is the amount of contextual information, not whether the misinformation was delivered via questions or a narrative, that determined whether RES occurred. Specifically, RES occurred when misinformation was presented in the context of the witnessed event narrative, but not when misinformation was presented as individual facts/questions. According to LaPaglia, presenting misinformation in an isolated manner increased its distinctiveness relative to embedding misinformation in a narrative, and this distinctive processing “maximized” encoding of the misinformation regardless of whether participants were tested previously – thereby removing the ability for testing to potentiate misinformation encoding. When this logic is applied to understand the aforementioned studies that reported a testing effect (LaPaglia & Chan, 2012; Memon et al., 2010b; Pansky & Tenenboim, 2011), it highlights the critical commonality across these studies – that participants encoded the misinformation in a distinctive manner.

**Subject population.** This is not strictly an encoding variable, but different subject populations (e.g., children, older adults) often encode materials differently. In one study (Chan et al., 2009), younger and older adults showed similar levels of RES for false recall of the
misinformation. In another study (LaPaglia, Wilford, Rivard, Chan, & Fisher, 2014), primarily Caucasian college students and primarily Hispanic college students demonstrated similar magnitudes of RES. Lastly, a recent study (Brackmann, Otgaar, Sauerland, & Howe, 2016) showed that children may be more susceptible to RES than young adults.

**Storage variables.** Variation in the retention interval between encoding and retrieval is one of the most straightforward ways for researchers to directly target memory storage. The influence of retention interval on RES is complex, given that four major events exist in its design: a) witnessed event, b) retrieval practice, c) misinformation presentation, d) final test. The retention interval between any of these consecutive time points have been manipulated in various studies.

**Delay between witnessed event and retrieval practice.** In most studies (e.g., Chan et al., 2009), the delay is quite short (often immediate), although Chan and LaPaglia (2013) inserted a substantial (i.e., > 24 hr) delay before retrieval practice in two of their experiments (see also Experiment 2 in the present study, to be described later). Based on the results of these experiment, RES can occur regardless of whether retrieval practice is delayed or not.

**Delay between retrieval practice and misinformation presentation.** This delay has been manipulated in a few studies. Although most studies presented the misinformation within 1 hr of the witnessed event, several studies showed that RES persisted when presentation of the misinformation was delayed by 24 hr to a week (Brackmann et al., 2016; Chan & Langley, 2011; Chan & LaPaglia, 2011). However, in a modified free choice recognition test in which the misinformation never appeared as a response candidate, Chan and LaPaglia (2013) showed that delaying the misinformation by 48 hr eliminated the RES effect for recognition memory of the witnessed event detail. Specifically, misinformation did not reduce the correct recognition rate regardless of whether participants had received an initial test. Although these findings appear inconsistent at first blush, a closer examination suggests otherwise. In Chan and Langley (and
also Chan & LaPaglia, 2011), the RES effect was apparent for misinformation production; however, similar to Chan and LaPaglia (2013), delayed presentation of the misinformation did not impair correct recall. These findings are consistent with the reconsolidation hypothesis, which suggests that only misinformation encountered shortly after retrieval interferes with reconsolidation of the witnessed event memory.

**Delay between misinformation and final test.** Most studies used a short delay (within 30 min), but reliable RES has also been reported when the final test occurred up to a week after participants were exposed to the misinformation (Chan & Langley, 2011; Chan & LaPaglia, 2013).

**Retrieval variables.** Several factors affect performance at the stage of retrieval. In the RES procedure, there are at least two retrieval events; the first is a retrieval practice/initial test phase that occurs before participants encounter misinformation, and the second is the final test, which occurs after participants encounter misinformation.

**Format of retrieval practice.** Most studies used cued recall (i.e., participants were asked to recall a specific detail about the witnessed event) for the initial test, and these studies have typically demonstrated an RES effect. Several studies have examined the influence of prior retrieval on subsequent suggestibility using other types of initial test, and the evidence here is more mixed. For example, Wilford et al. (2014) reported RES with a free recall initial test, whereas LaPaglia et al. (2014) showed RES with a Cognitive Interview initial test. In contrast, Holliday (2003a; 2003b) did not find an RES effect following the Cognitive Interview, and Gabbert et al. (2012) showed a testing effect following the Self-Administered version of the Cognitive Interview. Conspicuously absent is recognition. To our knowledge, no studies to date has investigated whether an initial recognition test would exacerbate subsequent eyewitness suggestibility.
Format of the final test. A variety of final test types have been examined, with the most common being cued recall, where a robust RES effect has been reported across a number of studies (e.g., Chan et al., 2009; Gordon et al., 2015). Chan, Wilford, and Hughes (2012) and Huff et al. (2013) used source recognition as the final test. Here, the results are mixed, with the former reporting an RES effect and the latter reporting a testing effect. But because these studies also differed on other methodological variables (e.g., Chan et al. used video materials for the witnessed event, item cued recall for the initial test, and a narrative for the misinformation that purportedly recapped the video, whereas Huff et al. used individual slides of household scenes for the original event, category cued recall for the initial test, and delivered the misinformation via what was purported to be recall sheets from other participants), it is not possible to ascertain why they produced different results. In addition, the RES effect has been demonstrated when the final test was forced choice recognition (but see LaPaglia & Chan, 2012 who reported a testing effect with a lineup identification task; Thomas et al., 2010), MMFR (Chan et al., 2009; Chan & LaPaglia, 2011; Gordon & Thomas, 2014), and narrative recall (Gordon & Thomas, 2017 in this test, participants were asked to recall what they remembered from the narrative rather than from the witnessed event). Both Gabbert and colleagues (2012) and Wilford et al. (2014) used free recall as the final test; here, Gabbert et al. found a testing effect, but Wilford et al. reported an RES effect. Recently, Rindal and colleagues (2016) showed that the RES effect is absent in the modified two-alternative forced choice (2AFC) recognition test that required participants to choose between the correct answer and a new foil (but not the misinformation).

Warning. A popular technique to reduce the misinformation effect is to issue a warning about the veracity of the materials through which misinformation is delivered (Blank & Launay, 2014). To our knowledge, only one published study has examined the influence of a post-misinformation warning on RES. Here, Thomas et al. (2010) showed that a warning at the time of the final test eliminated RES, regardless of whether the final test was cued recall or 2AFC forced choice recognition.
As the above review illustrates, RES is now a relatively well-established phenomenon. In the following, we describe two experiments that leveraged the RES paradigm to examine an important, but under-investigated, issue regarding eyewitness memory – whether inconsistencies in witness memory reports are indicative of inaccuracy.

**Investigating the Relation Between Eyewitness Inconsistency and Suggestibility in the Context of RES**

In the present study, we examined to what extent participant witnesses, when confronted with the idea that they had changed their responses across previous memory tests, were able to recover from their erroneous memory reports. Because eyewitnesses are often required to complete multiple interviews prior to their in-court appearance, concerns regarding the credibility of the witness accounts can arise when the witness changes his/her memory reports across interviews (e.g., when the witness recalls that the perpetrator was wearing a red shirt during one interview and then recalls that the perpetrator was wearing a gray shirt during a subsequent interview). Mock jurors judge inconsistent witnesses as less credible than consistent witnesses (Berman, Narby, & Cutler, 1995; Brewer & Burke, 2002; Pozzulo & Dempsey, 2009). Similarly, judges, lawyers, and police officers consider inconsistent witness as less credible according to survey studies (Fisher & Cutler, 1995; Strömwall, Strömwall, Granhag, & Granhag, 2003). Indeed, a primary purpose of cross-examination, at least in practice, is to induce a witness to change his/her statements, thereby making the witness appear unreliable or perhaps even deceptive (Glissan, 1991). Although cross-examination is sometimes regarded as a powerful truth-seeking tool (Stevens, 1992; Wigmore, 1974), laboratory research has shown that, rather than serving its role as a safeguard for the truth, cross-examination is often detrimental to the accuracy of eyewitness accounts due to its adversarial nature (Zajac, O'Neil, & Hayne, 2012). For example, Valentine and Maras (2011) showed that prior to cross-examination, participants who had been exposed to misleading suggestions from a co-witness were less accurate than control participants (i.e., a misinformation effect). However, cross-examination eliminated the
difference in accuracy between the two groups, because the control participants’ accuracy had dropped to the level of the misled participants (see also Segovia, Strange, & Takarangi, 2015; Zajac & Hayne, 2003).

Logically speaking, inconsistent responses should be less accurate than consistent ones. Although it is difficult to ascertain ground truth in criminal investigations, when a witness provides different answers to the same question across multiple occasions, it is likely that at least one of those answers is wrong. Of course, changes in memory reports can often be the result of an innocent mistake. For example, the eyewitness might have misunderstood a question during the prior interview, forgotten a previously remembered detail, or “recovered” a previously inaccessible memory (i.e., reminiscence). However, changes in memory reports can also occur because the witness is exposed to misinformation between interviews – which is what happens in the RES procedure. Given that misinformation can alter participants’ memory reports, it begs the question of whether participants can correct a previous mistake if given the chance to do so? Moreover, what role, if any, does prior testing have on participants’ ability to correct such mistakes?

Despite the widespread beliefs that inconsistent witnesses are unreliable witnesses, and the well-known finding that misinformation can cause changes in witness memory reports, to our knowledge, no studies to date has examined whether misinformation-induced inconsistencies are predictive of eyewitness accuracy. Several studies, however, have examined the relation between witness inconsistency and accuracy in the absence of external influences such as misinformation. Among these studies, the general conclusion is that, surprisingly, consistency in memory reports is not a valid indicator of memory accuracy at the individual level. That is, witnesses who have provided fewer inconsistent responses across interviews are no more accurate than witnesses who have provided more inconsistent responses (Brewer, Potter, Fisher, Bond, & Luszcz, 1999; Fisher & Cutler, 1995; Fisher, Vrij, & Leins, 2013; Gilbert & Fisher, 2006).
The Present Experiments

In the present study, we sought to answer three questions. First, to what extent is misinformation-induced memory inconsistency predictive of subsequent accuracy? Second, can participants overcome the negative influence of misinformation if they were told that they had provided inconsistent responses previously? Third, what role, if any, does initial testing have on the ability for participants to correct a former mistake? The RES methodology offers a unique opportunity to investigate misinformation-induced memory inconsistencies. Because participants are exposed to misinformation between repeated interviews in this procedure, it allows us to observe the relation between both spontaneous and misinformation-induced inconsistencies and subsequent eyewitness memory accuracy.

The current experiments used a five-phase design (see Figure 2). During Phase 1, participants viewed a video of the target event. During Phase 2, participants either completed a cued recall test for the target event (the tested participants) or not (the control participants). During Phase 3, all participants listened to a postevent narrative that contained some misleading information. During Phase 4, all participants took a final cued recall test and then completed a series of distractor tasks for a total of 75 min. During Phase 5, participants received a 24-question 2AFC recognition test. Each question in this test included the correct answer and a foil as response options. For the 12 items on which participants were misled, the foil matched the misinformation. All participants were given in-person, verbal instructions by the experimenter that the recognition test was their final memory test and that they needed to make sure that they pick the correct answer. In addition, the tested participants were also warned that they had sometimes given different answers to the same questions during the initial (Phase 2) and final (Phase 4) recall tests. These instructions were given to participants as an analogue of the type of questions that witnesses sometimes face when a prior inconsistency was identified (e.g., “you
told me different things about this on previous occasions, so which is it?").\(^1\) We hypothesized that telling participants that they had provided inconsistent answers might serve as a warning, and participants might be able to use this warning to reduce the impact of misinformation. Post-warning (i.e., warning provided to participants after they have encoded the misinformation, Blank & Launay, 2014) has been demonstrated to eliminate the RES effect (Thomas et al., 2010). However, such targeted warnings (e.g., “we cannot verify the accuracy of the narrative”) are difficult to implement in real-world settings, because they require the person who administers the warning to know the source of the misinformation. Inconsistency-based warnings, however, do not suffer from this shortcoming, because one can issue these warnings without knowing whether the witness has been exposed to misinformation.

Each experiment was conducted in two sessions scheduled one week apart. In Experiment 1, the one-week delay followed Phase 2 (i.e., the initial recall test for the tested participants, see the top half of Figure 2). In Experiment 2, the one-week delay followed Phase 1 (i.e., immediately after the witnessed event, see the bottom half of Figure 2). The two experiments were otherwise identical. Note that in both experiments, all participants took the Phase 4 final recall test and the Phase 5 final recognition test at the same retention interval following the witnessed event. We selected these specific delays to provide a contrast between a situation in which initial testing should have a large (Experiment 1) or negligible (Experiment 2) benefit on retention of the witnessed event. Specifically, in Experiment 1, the initial test occurred

\[\text{\footnotesize 1 We realize that our procedure (including both the instructions and the recognition test) did not provide a close approximation of the type of scenarios that actual witnesses may encounter in real-life. For example, eyewitnesses do not normally answer 2AFC recognition questions with the correct answer and a foil as response options. The purpose of the present paradigm was to examine participants’ memory accuracy after they have committed memory inconsistencies.}\]
soon after encoding. Performing retrieval practice at this time should help preserve retention and slow forgetting of the target event details. In contrast, in Experiment 2, the initial test happened one-week after encoding, by which time forgetting should have already reached asymptote, and performing retrieval practice at this point would likely provide little to no benefit for retention of the witnessed event details (Spitzer, 1939). Altogether, the timing parameters of Experiment 1 was expected to produce a situation in which retrieval practice would lead to a substantial gain for memories of the witnessed event memories relative to control (similar to the situation depicted in the right side of Figure 1), whereas the timing parameters of Experiment 2 should weaken, or perhaps eliminate, the ability for retrieval practice to enhance memories of the witnessed event relative to control (see the left side of Figure 1). Consequently, according to the TPL account, the RES effect should be smaller in Experiment 1 than in Experiment 2. Moreover, given that a stronger benefit of testing should emerge in Experiment 1 than Experiment 2, we expect that the inconsistency-warning would also have a greater benefit in Experiment 1.

Experiment 1

Method

Participants. Participants were 74 undergraduate students from a large Midwestern University who completed the experiment in exchange for course credit. The data for five participants were removed from final analysis because English was not their primary language. The data from an additional five participants were removed due to an experimenter error. Therefore, data from 64 participants were included in the final analyses. There were 33 participants in the tested condition and 31 in the control condition.

Materials and procedure. Participants consented to participate in a two-part study with Session 2 scheduled exactly one week after Session 1. Participants were tested in groups of up to eight while seated in front of computers separated by dividers. First, participants viewed a 43-minute video depicting a terrorist plot (Chan et al., 2009). Next, participants in the control
condition were dismissed, whereas those in the tested condition took an immediate, 18-question cued recall test. Each question targeted a different critical detail (e.g., what did the terrorist use to knock out the flight attendant?). Participants were told to answer as many questions as possible while maximizing their accuracy, but they were not forced to produce a response. They had up to 20 s to answer each question but could advance to the next question by pressing the "Enter” key.

The tested participants were dismissed after they have completed this initial recall test. One week later, all participants returned and listened to an 8-minute narrative. The narrative presented a verbal summary of the witnessed event video. Six critical details were re-presented correctly in the narrative (e.g., the terrorist knocks the flight attendant unconscious with a hypodermic needle) – consistent items. Six critical details were not mentioned in the narrative (e.g., the terrorist knocks the flight attendant unconscious) – neutral items. Twelve critical details were presented incorrectly in the narrative (e.g., the terrorist knocks the flight attendant unconscious with a chloroform rag) – misled items. Among the 12 misled items, six were questioned during the initial recall test for the tested participants (misled-tested), and six were not questioned during the initial recall test (misled-nontested). Assignment of each detail to item type was counterbalanced across participants using three versions of the narrative. It is important to clarify here that the narratives were 1,115 words long on average, and we embedded the 12 pieces of incorrect details into each counterbalanced version of the narrative. Consequently, each participant listened to a narrative that presented mostly correct information, but with 12 pieces of misleading details.

After the narrative, all participants took a 24-question cued recall test, with each question targeting one critical detail (they had up to 20 s to answer each question). Participants were instructed both verbally by the experimenter and through written instructions on the computer screen that they should answer the questions based on their memory of the video.

After the final recall test, participants completed a series of distractor tasks for a total of 75 min. First, they completed a 25 min comprehension test in which they read several short
passages and answered questions. Next, they completed the computerized operation span task and the reading span task (Unsworth, Heitz, Schrock, & Engle, 2005), and then they played the videogame Tetris for a total of 50 min.

Lastly, participants were given 10 min to take a final 2AFC recognition test for all 24 critical details. Each question included the correct answer and a foil as response options, and the identity of the foil matched the misinformation for the 12 misled items. After the 75 min distractor phase and just before this final recognition test, the experimenter re-entered the experiment room and told participants that they would now take a final memory test, and that they should “make sure that [they] pick the correct answer.” Participants were also asked to indicate their confidence for each question.² The tested participants were given the following additional set of instructions:

“When you were doing the comprehension task, I took a look at your data over the network and compared your answers from the first test you took a week ago with the test you took earlier today. I noticed that you sometimes gave different answers to the same question. This is normal and people often remember different things over time. In this next part of the experiment, I’d like you to take the test one last time. Unlike the previous tests, this will be a multiple choice test with two choices per question. I’d like you to make sure that you pick the correct

² After data collection was completed, we discovered that the experimenter provided partially inaccurate instructions regarding the confidence ratings. Specifically, although the experimenter indicated that participants were to provide confidence using a 0-10 confidence scale, the recognition test actually featured a 1-7 confidence scale. After answering each question, participants were asked to type a number between 1 and 7 to indicate their level of confidence, with the words “random guess” underneath 1 and “positive” under 7. Moreover, the computer would only accept an integer response in the 1 – 7 range.
answer. After answering each question, I’d also like you to indicate your confidence on your answer…”

In reality, the experimenter never examined participants’ responses prior to giving these instructions. Importantly, across both experiments, all participants had changed their responses to at least one question – thus validating these instructions. Once participants finished the final recognition test, they were given a short survey on demographics, then debriefed and dismissed.

Results

Recalled details were coded into the categories of correct, misinformation (i.e., when participants responded to a question with the misinformation), other (i.e., when participants produced a response that neither matched the witnessed event detail nor the misinformation), and no answer (including responses similar to “I don’t know”). For both experiments, we first present results for the final recall test (Phase 4) and then the final recognition test (Phase 5). Lastly, we present results on the final recognition test depending on whether participants had previously provided inconsistent responses to the same questions – these results are, of course, only available for the tested participants (because they were the only participants tested twice prior to the final recognition test). For the sake of completeness, results for the initial (Phase 2) recall test are shown in Table 1, but they will not be discussed further.

Final recall performance. All recall probabilities are shown in Table 2. Because of the vast number of item and response combinations available, we conducted only theoretically-motivated planned comparisons. In particular, we examined the influence of initial testing (i.e., tested participants vs. control participants) on subsequent correct and misinformation recall probabilities. For correct recall, we examined the effects of initial testing for both the neutral items and the misled items. For misinformation recall, we examined the effects of initial testing for only the misled items (because only these items were contradicted by misinformation, and participants rarely spontaneously produce an incorrect answer that matched the misinformation
for neutral items). Unless otherwise noted, misled items referred to the average performance of the 12 misled items for the control participants and the six misled-tested items for the tested participants. We considered data for the six misled-nontested items when examining whether prior testing increased eyewitness suggestibility to only items that appeared on the initial test?

The major results are shown in the left side of Figure 3. For correct recall, a 2(between-subjects: tested participants vs. control participants) X 2(within-subjects: neutral items vs. misled items) ANOVA yielded significant main effects for initial testing, \( F(1, 62) = 28.71, p < .01, \eta_p^2 = .32 \), item type, \( F(1, 62) = 37.37, p < .01, \eta_p^2 = .38 \), and the interaction between these variables, \( F(1, 62) = 4.12, p = .05, \eta_p^2 = .06 \). The main effects showed that, overall, the tested participants (\( M = .33 \)) outperformed the control participants (\( M = .15 \)) – i.e., there was a testing effect; moreover, participants were more likely to recall the correct event details for neutral items (\( M = .33 \)) than for misled items (\( M = .15 \)) – i.e., there was a misinformation effect. Perhaps more importantly, the interaction showed that presentation of the misinformation reduced the beneficial effects of testing by about half. Without misinformation (i.e., neutral items), participants who completed the immediate recall test were far more likely to recall the correct details one week later (\( M = .45 \)) than the control participants (\( M = .22 \)). This beneficial effect of testing was substantially reduced for details that were contradicted by misinformation; however, the tested participants (\( M = .21 \)) still outperformed their control counterparts (\( M = .09 \)). Another interpretation of this interaction is that the misinformation effect, as estimated by the reduction of correct recall probability between the neutral and misled items, was exacerbated by initial testing. Specifically, the reduction in correct recall due to misinformation was .13 (\( d = .51 \)) for the nontested participants, but .24 for the tested participants (\( d = 1.01 \)). It is important to note that this interaction must be interpreted with caution. Because control participants’ correct recall rate for the neutral items were closer to floor level (\( M = .21 \)), misinformation would likely have less influence on these participants than on the tested participants.
For misinformation recall, an independent samples t-test showed that the between-subjects RES effect was not significant, \( t(62) = 1.07, p = .29, d = 0.27 \), even though the tested participants (\( M = .55 \)) were slightly more likely to recall the misinformation than the control participants (\( M = .49 \)). Next, we compared the proportion of misinformation recall between items that did (\( M_{\text{misled-tested}} = .55 \)) and did not (\( M_{\text{misled-nontested}} = .51 \)) appear during the initial test for the tested participants. Here, too, the influence of RES at a within-subjects level was not significant, \( t(32) = 0.77, p = .44, d = 0.14 \). Overall, as we had predicted, Experiment 1 demonstrated that retrieval practice, when implemented soon after the occurrence of a witness event, can enhance delayed eyewitness memory performance (as shown by the higher correct recall rate for the tested participants relative to the control participants). However, despite this advantage, testing did not inoculate eyewitnesses from the negative impact of misinformation.

**Final recognition accuracy and confidence.** Similar to how we analyzed the final recall results, a 2 (between-subjects: tested vs. control) X 2 (within-subjects: neutral vs. misled) ANOVA for correct recognition showed a powerful main effect for item type, \( F(1, 62) = 110.00, p < .01, \eta_p^2 = .64 \), but neither the main effect of initial testing nor the interaction was significant, \( Fs < 1.94, ps > .16 \) (see the left side of Figure 4 for means). The main effect of item type indicates a misinformation effect – participants were far more likely to choose the correct option in the 2AFC test for a neutral item (\( M = .78 \)) than for a misled item (\( M = .39 \)). Planned comparisons showed that initial testing did not affect accurate recognition for either neutral items (\( M_{\text{control}} = .76, M_{\text{tested}} = .79 \)) or misled items (\( M_{\text{control}} = .35, M_{\text{tested}} = .42 \)), both \( ts < 1.21, ps > .23 \). These results are notable for two reasons. First, for the neutral items, initial testing produced almost no benefits for accurate recognition even though it had a substantial benefit for recall, a pattern that replicates previous research (e.g., Chan & McDermott, 2007; Jones & Roediger, 1995). Second, prior to conducting this experiment, we believe that telling the tested participants that they had produced inconsistent responses during previous tests might serve as a warning, and that participants might leverage the benefits of initial testing to overcome the negative
influence of misinformation or RES (cf., Thomas, Bulevich, & Chan, 2010). The present data, however, showed that recent exposure to misinformation had an overwhelming influence on recognition of a week-old witnessed event, regardless of whether participants had received retrieval practice or not.

We now present results regarding participants’ confidence in the final recognition test. Here, we examined the mean confidence associated with recognition of the correct detail and the misinformation in a 2(between-subjects: tested vs. control) X 2(within-subjects: neutral vs. misled) X 2(within-subjects: chose the correct option vs. chose the misinformation) ANOVA (see the left side of Figure 5). Both within-subject variables produced a main effect, $F$s $> 27.59$, $ps < .01$, $\eta_p^2$s $> .39$. The main effect of item type indicates that participants provided higher confidence ratings for misled items ($M = 5.65$) than for neutral items ($M = 4.54$), whereas the main effect of response type shows that participants provided higher confidence ratings when they chose the correct detail ($M = 5.51$) than when they chose the misinformation ($M = 4.69$). Perhaps most importantly, item type interacted with response type, $F(1, 42) = 10.13, p < .01, \eta_p^2 = .19$. This interaction showed that, for neutral items, incorrect responses (i.e., when participants chose the foil) were associated with lower confidence ratings ($M = 3.87$) than correct responses ($M = 5.21$) – i.e., confidence was, on average, diagnostic of accuracy. However, for misled items, this difference was eliminated, such that participants provided similar confidence ratings for both correct ($M = 5.81$) and incorrect responses (i.e., when they chose the misinformation instead of the correct answer, $M = 5.50$). No other main effects or interactions were significant. Hence, for both the tested participants and their control counterparts, presentation of the misinformation eliminated the diagnosticity of confidence ratings.

**Relation between response inconsistency and final recognition performance.** We now examine whether the level of response inconsistencies was predictive of performance on the final recognition test. To achieve this, we examine the proportion of neutral and misled questions
for which participants had provided different responses between the initial and final recall tests.\(^3\)

As expected, participants were less likely to provide inconsistent responses for neutral items (\(M = .41\)) than for misled items (\(M = .68\)), \(t(32) = 6.17, p < .01, d = 1.07\). More importantly, we examined whether participants who changed their responses more frequently were also more likely to succumb to misinformation during the final recognition test — after they were informed that they had provided inconsistent responses previously. Here, we examined the correlation between the proportion of questions with inconsistent responses and the participant’s final recognition accuracy. Specifically, response inconsistency for neutral items (between the initial and final recall test) was predictive of final recognition accuracy for these items, \(r = -.39, p = .02\), see the top left quadrant of Figure 6. Similarly, response inconsistency for misled items was also predictive of final recognition accuracy for these items, \(r = -.46, p < .01\), see the top right quadrant of Figure 6. Both of these negative correlations indicate that the less consistent participants were also more likely to select the incorrect response during the final recognition test.

The above data illustrate the relation between response inconsistency and final recognition accuracy at the individual/participant level. We also examined this relation at the item level. Specifically, was final recognition performance dependent upon whether participants had provided consistent vs. inconsistent responses for that item previously? A 2(neutral items, misled items) X 2(consistent, inconsistent) within-subjects ANOVA showed that both main effects and the interaction were significant, \(F$s > 21.24, ps < .01, \eta^2_p$s > .41\). The interaction

\(^3\) For this analysis, we considered a response as having changed if the participant provided meaningfully different responses across the two recall tests. Minor wording differences in answers such as “she was using a computer” and “she was on the computer” were not considered inconsistent responses.
indicates that final recognition accuracy was substantially lower for misled items that had elicited inconsistent responses ($M = .26$) than those that had elicited consistent responses ($M = .77$), $t(31) = 7.16, p < .01, d = 1.27$. This pattern, however, was far more muted for the neutral items. Here, recognition accuracy for items that had elicited inconsistent responses ($M = .73$) was still lower than the items that elicited consistent responses ($M = .86$), $t(31) = 2.24, p = .03, d = 0.39$, but the difference was much smaller than that exhibited by the misled items.

**Experiment 2**

**Method**

**Participants.** Participants were 66 undergraduate students from a large Midwestern University who completed the experiment in exchange for course credit. Data from six participants were excluded from analyses because English was not their primary language. Therefore, data from 60 participants was used in the final analyses, with 29 participants in the control condition and 31 in the tested condition.

**Materials and procedure.** All materials and procedure were identical to Experiment 1 except that participants were dismissed for one week after they watched the target video. After this retention interval, participants in the tested condition took the first cued recall test before they listened to the audio narrative, whereas participants in the control condition listened to the narrative immediately upon their return (see bottom of Figure 2). The verbal instructions that were given to the tested participants just before the final recognition test was adapted to match the scheduling of Experiment 2.

**Results**

**Final recall performance.** The main findings of interest are shown on the right side of Figure 3. For correct recall, a 2(tested vs. control) X 2(neutral vs. misled) ANOVA revealed a marginally significant main effect for initial testing, $F(1, 58) = 3.69, p < .06, \eta_p^2 = .06$. Unlike
Experiment 1, when initial testing increased correct recall, the marginal main effect of testing here was driven by a reduction in correct recall for the tested participants ($M = .13$) relative to the control participants ($M = 17$). Indeed, as we had predicted, delaying initial testing by a week eliminated the beneficial effects of retrieval practice on even the neutral items, ($M_{tested} = .19$ and $M_{control} = .20$), $t < 1$. There was also a significant main effect for item type, $F(1, 58) = 9.81, p < .01, \eta^2_p = .15$. Unsurprisingly, misinformation reduced the likelihood of correct recall ($M_{misled} = .10$ and $M_{neutral} = .19$). The interaction between initial testing and item type was not significant, $F(1, 58) = 1.15, p = .29, \eta^2_p = .02$.

Data for misinformation recall showed a powerful between-subjects RES effect, $t(58) = 7.00, p < .01, d = 1.81$, with the tested participants recalling the misinformation at an alarming rate ($M = .70$) relative to the control participants ($M = .39$). Moreover, among the tested participants, the misinformation recall rate was higher for items questioned on the initial test ($M_{misled-tested} = .70$) than for items omitted from the initial test ($M_{misled-nontested} = .61$), $t(30) = 2.32, p = .03, d = 0.42$ — a within-subjects RES effect. Lastly, initial testing appeared to have increased participants’ susceptibility to misinformation as a whole, as the tested participants were also more likely to recall misinformation for previously nontested details ($M_{misled-nontested} = .61$) relative to the control participants ($M = .39$), $t(58) = 5.10, p < .01, d = 1.32$. Altogether, these results demonstrate that RES might not be confined to just the event details that were explicitly tested (see also Gordon & Thomas, 2014; Wilford et al., 2014).

**Final recognition accuracy and confidence.** A 2(tested vs. control) X 2(neutral vs. misled) mixed ANOVA for the probability of final correct recognition yielded no main effect for initial testing, $F(1, 58) = 1.08, p = .30$, a main effect for item type, $F(1, 58) = 103.55, p < .01, \eta^2_p = .64$, and most importantly, a crossover interaction between these variables (see the right side of Figure 3), $F(1, 58) = 22.58, p < .01, \eta^2_p = .28$. Unlike the results from the final recall test (in which initial testing failed to boost correct recall for either neutral or misled items), initial testing — even after a one-week delay — increased correct recognition of the neutral items ($M_{control}$
participants = .61, $M_{\text{tested participants}} = .73), t(58) = 2.09, p = .04, d = 0.54, but it reduced correct recognition for the misled items ($M_{\text{control participants}} = .43, M_{\text{tested participants}} = .24), t(58) = 4.54, p < .01, d = 1.17. In other words, the tested participants were far more likely to choose the misinformation instead of the correct detail in a forced-choice test than the control participants, even after the tested participants were told that they had provided inconsistent responses previously.

We now examine the confidence associated with participants’ recognition responses. Similar to Experiment 1, we conducted a 2(tested vs. control participants) X 2(neutral vs. misled items) X 2(correct vs. misinformation recognition) mixed ANOVA (see the right side of Figure 4). There was a main effect of item type, $F(1, 40) = 14.21, p < .01, \eta_p^2 = .26, such that participants provided higher confidence for the misled items ($M = 5.32$) than for the neutral items ($M = 4.75$) — a pattern that replicated the results from Experiment 1. The interaction between item type and response type was also significant, $F(1, 40) = 33.34, p < .01, \eta_p^2 = .46$. Similar to Experiment 1, confidence ratings were diagnostic of accuracy for the neutral items, such that participants provided lower confidence ratings for incorrect responses ($M = 4.33$) than for correct responses ($M = 5.18$). However, the opposite pattern occurred for the misled items, such that participants indicated higher confidence for incorrect responses ($M = 5.87$) than for correct responses ($M = 4.77$). Perhaps most importantly, the three-way interaction was significant, $F(1, 40) = 9.59, p < .01, \eta_p^2 = .19$. Specifically, confidence results for the control participants mirrored the pattern exhibited in Experiment 1, such that confidence ratings were diagnostic of accuracy for the neutral items ($M_{\text{correct}} = 5.04, M_{\text{foil}} = 4.35$) but not for the misled items ($M_{\text{correct}} = 5.10, M_{\text{foil}} = 5.31$). However, a pronounced, crossover pattern emerged for the tested participants. Whereas correct responses for neutral items were associated with higher confidence than incorrect responses ($M_{\text{correct}} = 5.31, M_{\text{foil}} = 4.30$), the opposite was true for misled items ($M_{\text{correct}} = 4.44, M_{\text{foil}} = 6.43$). Therefore, not only did initial testing increased the likelihood that participants
would select the incorrect, misleading alternative in the forced choice recognition test, it also increased the confidence associated with these incorrect responses.

**Relation between response inconsistency and final recognition performance.** We now turn to results regarding response inconsistencies. As expected, participants were far more likely to change their answers across the recall tests for misled items ($M = .77$) than for neutral items ($M = .43$), $t(30) = 8.96, p < .01, d = 1.61$. Moreover, similar to the results of Experiment 1, the level of response inconsistency for the misled items was predictive of final recognition accuracy for these items (see the right side of Figure 5), $r = -.61, p < .01$. However, unlike Experiment 1, the same correlation was not significant for the neutral items, $r = -.23, p = .22$.

To examine the relation between response inconsistency and final recognition performance at the item level, we conducted a 2(neutral items, misled items) X 2(consistent, inconsistent) within-subjects ANOVA. The results of this analysis were consistent with those in Experiment 1, with both the main effects and the interaction reaching significance, $F$s $> 6.83, ps < .02, \eta^2$s $> .23$. Critically, the interaction shows that, once again, participants were much less likely to choose the correct response alternative during the final recognition test for the *misled items* that had previously elicited inconsistent responses ($M = .15$) than those that had elicited consistent responses ($M = .50$), $t(22) = 4.04, p < .01, d = 0.84$, but this difference was virtually absent for the *neutral items* ($M_{inconsistent} = .76, M_{consistent} = .75$), $t(30) = 0.96, p = .32, d = 0.04$.

**General Discussions**

In two experiments, participants were exposed to misinformation a week after they viewed the witnessed event. Half of the participants in Experiment 1 completed their initial “interview” immediately after the witnessed event, whereas half of the participants in Experiment 2 completed their initial “interview” a week after the witnessed event. As predicted, a substantial memorial benefit of testing was observed in Experiment 1 – for event details that were not contradicted by misinformation. This benefit of testing was reduced considerably for
misled items – although it was still reliable. Notably, prior retrieval did not significantly increase reporting of misinformation in the final recall test of Experiment 1, although it also did not reduce reporting of the misinformation (i.e., no RES effect for misinformation recall was found). The results in Experiment 2, however, differed dramatically from those in Experiment 1. Not only did initial testing fail to increase correct recall of the neutral items, it increased reporting of misinformation for the misled items in the final recall test – i.e., an RES effect.

In the present experiments, participants also took a final 2AFC recognition test, and prior to this test, the tested participants were informed explicitly that they had provided different answers to the same question previously. We reasoned that this warning might help participants to overcome the influence of RES when they were given an extra, final opportunity to answer questions about the witnessed event – during which the correct answer was provided as one of the response options. Clearly, our data do not support this hypothesis. We now consider some of the more noteworthy findings across these two experiments.

In Experiment 1, although the immediate initial test significantly boosted accurate recall of the neutral items during the recall test one week later, it did not increase correct recognition of these items. However, in Experiment 2, although the delayed initial test did not increase accurate recall of the neutral items, it did enhance accurate recognition of these items. If this finding is replicable, then it shows that even delayed initial testing might, under some circumstances, benefit subsequent memory performance. Given the novel and unexpected nature of this finding, any attempts at providing an explanation would be necessarily speculative. Nonetheless, we highlight this result here because it may prove a fruitful avenue for future investigation.

Perhaps most importantly, 2AFC recognition performance of the misled items mirrored those from the final recall test. Therefore, the same pattern of results persisted from the final recall test to the final recognition test, thus demonstrating that participants were unable to
overcome the influence of RES with the inconsistencies warning. This finding is somewhat surprising when compared to the results from another warning study (Thomas et al., 2010), which showed that warning subjects that the postevent narrative may contain inaccuracies eliminated the RES effect. There are several differences between the implementation of warnings in the present study relative to the Thomas et al. study, and any of these differences could have contributed to the different results. First, Thomas et al. used a targeted warning that informed participants that the narrative might contain misinformation (a warning that would be difficult to implement in real-life), whereas the inconsistencies warning used here did not explicitly inform subjects of potential inaccuracies in the postevent narrative. Second, in Thomas et al., the warning was delivered just before the second/final recall test; however, in the present experiments, the inconsistencies warning was delivered after participants had completed the second/final recall test, but before the final recognition test. This difference is important, as the final recall test gave participants an opportunity to commit the misinformation-induced errors before the warning was issued (Schreiber & Sergent, 1998). Lastly, Thomas et al. conducted their experiment in a single experimental session, but the current experiments had a one-week interval between presentation of the witnessed event and the final test.

A potential concern with the current experimental design is that we did not include a condition in which the tested participants were not given the inconsistencies warning. It is therefore possible that the inconsistencies warning was beneficial, and that the tested participants would have performed even worse on the final recognition test had they not been given this warning. In other words, without this condition, we do not know if the inconsistency warning was ineffective at reducing false recognition, or if the warning was indeed effective, but the current experiment simply lacked the necessary baseline condition to detect the benefits of such a warning. However, a closer examination of the recognition performance in Experiment 2 suggests that the latter possibility is unlikely, given that the tested participants selected the misinformation 76% of the time and provided near-ceiling level of confidence for these errors. It
is hard to imagine that participants could have done much worse from a suggestibility perspective.

Our results show that committing a misinformation-induced error had a relatively long-lasting consequence on memory (e.g., Schooler, Foster, & Loftus, 1988). Furthermore, when participants chose the misinformation option during the final recognition test, they provided confidence ratings that were comparable to (or exceeded) those for the correct option. These results highlight the fact that not all mistakes are created equal—although spontaneous memory errors (i.e., when participants picked the misinformation option for a neutral item) were associated with lower confidence than accurate responses (see also Wixted, Mickes, Clark, Gronlund, & Roediger, 2015), misinformation-induced errors were not. In our opinion, given that it is often difficult to know whether and to what misinformation an eyewitness might have been exposed (and thus difficult to tell whether an inconsistent response might have been driven by misinformation), it remains risky to use eyewitness confidence as an indicator of accuracy overall (Fox & Walters, 1986; Penrod & Cutler, 1995; Shaw, McClure, & Dykstra, 2007), and even more so when court/police records indicate that a witness might have been exposed to misinformation (“Prosecutors Appeal Overturned Conviction of ‘Making a Murderer’ Subject Brendan Dassey,” n.d.).

A major interest of this study was to examine the relation between response inconsistencies and subsequent memory accuracy. As stated previously, the RES methodology provides a suitable environment to investigate this question, because participants are exposed to multiple “interviews” and are exposed to misinformation for some of the witnessed details. At the item level, it is perhaps not surprising that participants would exhibit poorer recognition performance for the items that had elicited inconsistent responses than those that had elicited consistent responses—participants might have provided inconsistent responses for a particular item because they had a weak memory for that item. Therefore, this finding may simply reflect an item-selection artifact. More important, however, was that the relation between response
inconsistency and later recognition accuracy differed depending on item type, such that participants were far more likely to err during the final recognition test for the misled items than the neutral items, even when we confined the analysis to only items that had elicited inconsistent responses ($M_{\text{neutral}} = .75$ vs. $M_{\text{misled}} = .21$ averaged across the two experiments). The difference between spontaneous response inconsistencies (i.e., inconsistent responses for neutral items) and misinformation-induced response inconsistencies was also evident at the individual level. Here too, we found that response inconsistencies predicted accuracy during the final recognition test, but the association was stronger when the inconsistent responses were misinformation-driven rather than spontaneous (at least in Experiment 2). Together, these results show that not all memory errors are created equal.

The general public, including actors of the judiciary, believe that inconsistent witnesses are unreliable. Despite the importance and prevalence of this belief, scant empirical research has examined whether this assumption is justified, and the existing studies painted a picture whereby the relation between response inconsistencies and accuracy was weak at best. The present research, however, shows that the relation between inconsistency and accuracy can be quite strong – particularly for inconsistencies induced by misinformation. When combining the data across the two experiments, spontaneous inconsistencies (i.e., when participants provide inconsistent responses to a neutral item) was moderately predictive of final recognition accuracy for these items, accounting for roughly 10% of its variance, $r = -.31$, $F(1, 62) = 6.70, p = .01$. However, the predictive power for misinformation-induced inconsistencies for final recognition of these items was much stronger, accounting for roughly 30% of its variance, $r = -.55$, $F(1, 62) = 26.92, p < .01$.

Previously, Chan and LaPaglia (2011) argued that, theoretically, it is important to distinguish systematic from nonsystematic eyewitness errors because the former can be manipulated by system actors (e.g., an officer might knowingly introduce misinformation to obtain a conviction) and might be associated with greater confidence than the latter (Chan et al.,
In the present study, we further showed that these two types of errors, in the form of response inconsistencies, also held different predictive powers for subsequent memory accuracy. Distinguishing systematic from nonsystematic errors is straightforward in experimental settings, where the researcher can control the introduction of misinformation. However, in the real-world, it is often difficult to know if an eyewitness had been exposed to misinformation, let alone knowing whether particular inconsistent witness statements were influenced by misinformation. Therefore, the applied value of the current findings does not rest on the possibility of identifying the source of a memory inconsistency after witnesses have been exposed to misinformation; rather, these results emphasize the critical importance of preventing the unintentional introduction of misinformation by investigators (e.g., by training investigators to use interview techniques designed to minimize contamination, such as the Cognitive Interview, Fisher & Geiselman, 1992; R. E. Holliday, 2003b; Memon, Meissner, & Fraser, 2010a).

Concluding Remarks

Under many circumstances, retrieval practice is a powerful memory enhancer. Not only does retrieval strengthens subsequent recall of the tested information (Roediger & Butler, 2011), it also potentiates later learning of new information (Pastotter & Bauml, 2014). Research on the RES effect, however, has highlighted some conditions under which prior retrieval can be detrimental to memory performance, and theoretical work in this literature has shown that the very mechanisms by which retrieval strengthens new learning might also contribute to RES. In the current study, we used the RES paradigm to examine whether eyewitness inconsistencies are indicative of accuracy, and we found that the answer to this question depends partly on whether the inconsistencies were evoked by external influences.
Author Contributions

Jason Chan conceptualized, designed, programmed the experiments, analyzed the data, and wrote the manuscript. Kathryn Lang collected all data and coded some of the data. Krista Manley re-coded all data, assisted in analyses, and provided feedback for the manuscript.
References


testing can enhance suggestibility. Presented at the biennial meeting of the Society for Applied Research in Memory and Cognition, Sydney, Australia.


Krawczyk, M. A. C., Fernandez, R. S., Pedreira, M. E., & Boccia, M. M. (2017). Toward a better
understanding on the role of prediction error on memory processes: From bench to clinic. *Neurobiology of Learning and Memory, 142*, 13–20.
LaPaglia, J. A. (2013). *Testing can enhance or reduce suggestibility: The importance of contextual detail during misinformation exposure*. Ames, IA.


Table 1

*Initial Recall Test Performance for Experiments 1 and 2*

<table>
<thead>
<tr>
<th>Response type</th>
<th>Correct</th>
<th>Misinformation</th>
<th>Other</th>
<th>No answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experiment 1</td>
<td>.49 (.12)</td>
<td>.06 (.05)</td>
<td>.36 (.14)</td>
<td>.08 (.12)</td>
</tr>
<tr>
<td>Experiment 2</td>
<td>.19 (.10)</td>
<td>.09 (.08)</td>
<td>.54 (.16)</td>
<td>.18 (.20)</td>
</tr>
</tbody>
</table>

*Note.* Standard deviations are presented in parentheses. Differentiation of item types (i.e., consistent, neutral, misled) occurred when participants listened to the audio narratives (and after the initial recall test). Therefore, we presented the data here collapsed across response types.
Table 2.

*Final Recall Performance as a Function of Initial Testing Condition, Item Type, and Response Type*

<table>
<thead>
<tr>
<th></th>
<th>Correct</th>
<th>Misinformation</th>
<th>Other</th>
<th>No answer</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Experiment 1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Tested participants</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Consistent items</td>
<td>.72 (.18)</td>
<td>.02 (.06)</td>
<td>.23 (.18)</td>
<td>.03 (.08)</td>
</tr>
<tr>
<td>Neutral items</td>
<td>.45 (.21)</td>
<td>.07 (.11)</td>
<td>.41 (.19)</td>
<td>.07 (.14)</td>
</tr>
<tr>
<td>Misled-tested items</td>
<td>.21 (.17)</td>
<td>.55 (.22)</td>
<td>.21 (.18)</td>
<td>.04 (.07)</td>
</tr>
<tr>
<td>Misled-nontested items</td>
<td>.12 (.16)</td>
<td>.51 (.25)</td>
<td>.31 (.24)</td>
<td>.07 (.12)</td>
</tr>
<tr>
<td><strong>Control participants</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Consistent items</td>
<td>.61 (.26)</td>
<td>.04 (.09)</td>
<td>.27 (.21)</td>
<td>.08 (.15)</td>
</tr>
<tr>
<td>Neutral items</td>
<td>.22 (.22)</td>
<td>.05 (.09)</td>
<td>.58 (.22)</td>
<td>.16 (.20)</td>
</tr>
<tr>
<td>Misled-nontested items</td>
<td>.09 (.09)</td>
<td>.49 (.21)</td>
<td>.32 (.16)</td>
<td>.10 (.11)</td>
</tr>
<tr>
<td><strong>Experiment 2</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Tested participants</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Consistent items</td>
<td>.78 (.17)</td>
<td>.03 (.06)</td>
<td>.15 (.15)</td>
<td>.04 (.10)</td>
</tr>
<tr>
<td>Neutral items</td>
<td>.18 (.14)</td>
<td>.09 (.11)</td>
<td>.56 (.24)</td>
<td>.17 (.21)</td>
</tr>
<tr>
<td>Misled-tested items</td>
<td>.06 (.09)</td>
<td>.70 (.18)</td>
<td>.19 (.14)</td>
<td>.04 (.09)</td>
</tr>
<tr>
<td>Misled-nontested items</td>
<td>.10 (.14)</td>
<td>.61 (.16)</td>
<td>.23 (.16)</td>
<td>.06 (.10)</td>
</tr>
<tr>
<td><strong>Control participants</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Consistent items</td>
<td>.55 (.22)</td>
<td>.06 (.12)</td>
<td>.30 (.20)</td>
<td>.10 (.16)</td>
</tr>
<tr>
<td>Neutral items</td>
<td>.20 (.20)</td>
<td>.09 (.10)</td>
<td>.54 (.27)</td>
<td>.17 (.21)</td>
</tr>
<tr>
<td>Misled-nontested items</td>
<td>.14 (.13)</td>
<td>.39 (.16)</td>
<td>.34 (.16)</td>
<td>.12 (.13)</td>
</tr>
</tbody>
</table>

*Note.* Standard deviations are presented in parentheses.
Figure 1. Relation between the memory strengthening effects of retrieval practice and whether RES is predicted by the test-potentiated learning (TPL) account. The left half of the figure illustrates a situation whereby retrieval practice enhances memory of the witnessed event to a smaller extent than it does for memory of the misinformation. According to the TPL account, RES should occur in this situation. The right half of the figure illustrates the opposite situation, whereby retrieval practice enhances memory for the witnessed event more than it does for memory of the misinformation, and RES should not occur in this situation according to the TPL account.
Figure 2. Design for Experiments 1 and 2.
Figure 3. Correct and misinformation recall probabilities for neutral and misled items in the final recall test. Misinformation recall refers erroneous responses whereby participants responded to a question with the misinformation. Error bars display descriptive .95 CI.
Figure 4. Correct recognition probabilities for neutral and misled items in the final 2AFC recognition test. Error bars display descriptive .95 CI.
Figure 5. Confidence associated with correct and misinformation recognition in the final 2AFC recognition test. Confidence ratings ranged from 1 – 7. Error bars display descriptive .95 CI.
Figure 6. Relation between response inconsistency (during initial and final recall) and accuracy during final recognition.