Political Behavior
https://doi.org/10.1007/s11109-020-09597-3

ORIGINAL PAPER

Fake Claims of Fake News: Political Misinformation, Warnings, and the Tainted Truth Effect

Melanie Freeze[1] · Mary Baumgartner[2] · Peter Bruno[3] · Jacob R. Gunderson[4] · Joshua Olin[5] · Morgan Quinn Ross[6] · Justine Szafran[7]

Abstract

Fact-checking and warnings of misinformation are increasingly salient and prevalent components of modern news media and political communications. While many warnings about political misinformation are valid and enable people to reject misleading information, the quality and validity of misinformation warnings can vary widely. Replicating and extending research from the fields of social cognition and forensic psychology, we find evidence that valid retrospective warnings of misleading news can help individuals discard erroneous information, although the corrections are weak. However, when informative news is wrongly labeled as inaccurate, these false warnings reduce the news' credibility. Invalid misinformation warnings taint the truth, lead individuals to discard authentic information, and impede political memory. As only a few studies on the tainted truth effect exist, our research helps to illuminate the less explored dark side of misinformation warnings. Our findings suggest general warnings of misinformation should be avoided as indiscriminate use can reduce the credibility of valid news sources and lead individuals to discard useful information.

Keywords Tainted truth effect · Fake news · Warnings · Misinformation

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11109-020-09597-3) contains supplementary material, which is available to authorized users.

✉ Melanie Freeze
mfreeze@carleton.edu

Extended author information available on the last page of the article

Published online: 05 February 2020

"O, what a tangled web we weave when first we practice to deceive."
Walter Scott, Marmion

Introduction

Warnings of misinformation are an increasingly common feature of American political communication. The spread of misleading news through social media platforms during the 2016 U.S. election season provoked widespread discussions of and warnings about political misinformation (Allcott and Gentzkow 2017; Frankovic 2016; Guess et al. 2018a, b; Nyhan 2019; Silverman 2016; Silverman et al. 2016; Silverman and Singer-Vine 2016). In the months prior to the 2016 general election, one in four Americans read a fact-checking article from a national fact-checking website (Guess et al. 2018b, p. 10). Fact-checking organization growth accelerated in the early 2000s, and the number of fact-checking outlets continues to increase in the U.S. and around the world (Graves 2016; Graves et al. 2016; Spivak 2010; Stencel 2019). Due to the increased salience of political misinformation and rise of fact-checking organizations, people often encounter warnings regarding misinformation, but the quality and veracity of these warnings can vary considerably. In this article, we evaluate how invalid warnings of misinformation can lead people to distrust the information's source, cause people to discard accurate information, and ultimately impede memory.

Valid warnings of misinformation tend to originate from professional third-party organizations, target information that is actually misleading, and reduce the spread and acceptance of misinformation. For example, FactCheck.org, PolitiFact, and the Washington Post's Fact Checker are all organizations that investigate the veracity of claims made by political figures and news organizations, operate year-round, and view themselves as a distinct professional cohort within journalism guided by rules and norms (Graves 2016). Warnings originating from these organizations tend to be precise and issued neutrally.[1] Other institutions, such as Facebook, also devote resources to counteract false news through critical changes to algorithms and various policies. Working to retain users' trust and confidence in their site, Facebook's warnings of misinformation often seek to correctly identify and reduce the spread of actual misinformation, although these efforts have recently excluded the direct speech of politicians (Funke 2019; Kang 2019; Mosseri 2017).[2]

[1] For examples of high quality misinformation warnings see Cook and Lewandowsky (2011) and Nyhan and Reifler (2012).

[2] It is arguable that the validity of Facebook and other organizations' fact-checking efforts also vary, especially in early stages of development. In the immediate months following the 2016 election, Facebook collaborated with fact-checking organizations, flagged false news as "Disputed," and warned people of the status before they attempted to share the article. These "Disputed" tags were later replaced by a policy in which people viewing popular links were instead shown a series of "Related Articles" that included both misinformation and third-party fact-checker articles (Allcott et al. 2019, Appendix 4). Since 2016, Facebook's general strategy has been to "remove, reduce, and inform" (Lyons 2018a), and the organization continues to update and revise their approach to misinformation, drawing on machine learning tools and expanding fact-check efforts to photos and videos (Lyons 2018b). However, Facebook has recently taken a more hands-off approach to claims or statements made by politicians on their Facebook Page, an ad, or their website. These statements are considered direct speech and ineligible for third-party fact checking program (Kang 2019).

Irrespective of the source of a warning, the main criterion of whether or not a warning is valid is if it correctly targets misinformation and efficiently counters the effects of misinformation.

In contrast, less valid or invalid misinformation warnings are biased and inefficient. First, warnings of misinformation are biased when they target factual information rather than misinformation. Bias may be inadvertent but some misinformation warnings are intentionally designed to discredit information. Strategic elites may issue warnings of misinformation against news that is factually correct but unfavorable. Recently, the term "fake news," has been used by politicians and pundits around the world to discount news reports and organizations they find disagreeable in order to control political news and shape public opinion (Tandoc Jr. et al. 2018; Wardle and Derakhshan 2017; Wong 2019).

Second, warnings of misinformation may be less valid because their effects are inefficient and imprecise. In the U.S., President Donald Trump frequently uses the term "fake news" in tweets referencing the mainstream news media, especially in reaction to critical coverage or investigative reporting (Sugars 2019). These and other warnings of misinformation employed by President Trump are often so broadly construed that they could potentially target both misleading and accurate news (Grynbaum 2019a, b). For example, on March 28, 2019, President Donald Trump wrote "The Fake News Media is going Crazy! They are suffering a major "breakdown," have ZERO credibility or respect, & must be thinking about going legit. I have learned to live with Fake News, which has never been more corrupt than it is right now. Someday, I will tell you the secret!"[3]

While clumsy warnings may be able to counter misinformation, they are less valid because they often incur high unintended casualties. For example, in contrast to warnings that identify specific misleading facts, Clayton et al. (2019) find that general warnings of misinformation shown to people before news exposure reduce the perceived accuracy of both real and false news headlines. Mistrust and rejection of news is beneficial when that news is misleading, but when the mistrust and rejection spills over to real news, the potential drawbacks of misinformation warnings become apparent.

Pennycook and Rand (2017) also uncover other drawbacks of misinformation warnings. An "implied truth effect" emerges when some, but not all, false stories are tagged as misinformation. Those false stories which *fail* to get tagged are considered validated and seen as *more* accurate. Even legitimate misinformation warnings, if not fully deployed, can enhance the effects of misinformation in the larger system. Sophisticated organizations seek to employ nuanced and specific fact-checking techniques, but less valid warnings of misinformation continue to be used by both political elites and in broad public conversations on misinformation and the news media. Consequently, it is very important that we continue to investigate both the positive and negative effects of misinformation warnings in the realm of news media and political communications.

[3] Accessed June 8, 2019 at https://twitter.com/realdonaldtrump/status/1111209625825640448.

In this study, we investigate the potentially negative side effects of invalid, retrospective[4] misinformation warnings. To do this, we replicate and expand a relatively understudied area of research traditionally applied to the area of eyewitness testimony in the field of social cognition. Specifically, we investigate the *tainted truth effect*, which proposes that misdirected warnings of post-event misinformation can disadvantage memory of an original event by discrediting factual information and causing it to be discarded at the time of memory assessment (Echterhoff et al. 2007; Szpitalak and Polczyk 2011).

Drawing on Szpitalak and Polczyk's (2011) study on the tainted truth effect, we replicate and extend their three primary research questions to a political context. We first ask, after viewing a political event, how does later exposure to information and misinformation in a news article describing the event alter individuals' memory and recognition of the details from the original event? Second, when individuals are retrospectively exposed to a valid warning that the news article contained misinformation, are they able to discard the misinformation and remember the correct original event information? Third, do people discard accurate data when exposed to an invalid warning of misinformation? While all three research questions work together to build a picture of individual memory and information processing, the third question regarding the potential drawbacks of misinformation warnings, formally referred to as the tainted truth effect, is the focus of our research. Finally, building on Szpitalak and Polczyk's three primary questions, we also consider the mechanisms and nuances of misinformation warnings, that is, how these warnings influence the credibility of the warning's target and the certainty of memory.

From these questions, we derive a series of particular expectations. First, in the absence of a misinformation warning, we expect that individuals' memories of the original event will be strongly influenced by a post-event description, that is, a related news article. Receiving misleading (or accurate) post-event descriptions in a news article will decrease (or increase) respondents' ability to recognize original event details.

Hypothesis 1a (Misinformation Effect) Exposure to misleading information in a post-event description is expected to reduce memory recognition of the original event.

Hypotheses 1b (Information Effect) Exposure to accurate information in the post-event description is expected to increase the memory recognition of the original event details.

Second, respondents who are exposed to misinformation in the news article but are later warned about misleading information should recognize original event

[4] Retrospective warnings are warnings presented to an individual after misinformation exposure [see Blank and Launay (2014) for a review]. Echterhoff et al. (2007) recommend research on retrospective warnings of misinformation. These scholars argue retrospective warnings are more likely to mirror real life situations given the difficulty in identifying misinformation and warning people prior to exposure.

details and misinformation better than respondents who were exposed to misinformation without a warning.

Hypothesis 2a (Warning and the Memory Performance) Exposure to a valid retrospective misinformation warning will increase the ability to correctly recognize original event details.

Hypothesis 2b (Warning and the Misinformation Recognition) Exposure to a valid retrospective misinformation warning will reduce the incorrect recognition of misinformation as original event information.

Third, warnings of misinformation are expected to taint *all* information that is associated with the news article. Therefore, misinformation warnings, even when completely invalid (in the case where no misinformation is in the post-event description), should lead individuals to also reject accurate information that is associated with the news article and result in reduced memory accuracy compared to individuals who are not warned.

Hypothesis 3 (Tainted Truth Effect) Exposure to an invalid or imprecise retrospective misinformation warning will reduce the ability to correctly recognize original event details.

Finally, we expect trust to be fundamentally damaged by misinformation warnings. First, when warned of misinformation, individuals should be less trusting of their own memory and feel more uncertain about their responses. Second, warnings of misinformation should erode trust in the origins of the information and should lead people to view the news source as less credible.

Hypothesis 4a (Warning and Information Uncertainty) Exposure to a misinformation warning will increase memory uncertainty.

Hypothesis 4b (Perceived Credibility) Exposure to a misinformation warning will reduce the perceived credibility of the post-event description that is targeted by the warning.

We find evidence that retrospective, invalid misinformation warnings taint news and lead individuals to view the news as less credible. Increased skepticism produced by invalid misinformation warnings leads individuals to discard information that was in fact accurate, as predicted by the tainted truth hypothesis, and these invalid warnings are also associated with more memory uncertainty. In addition to the tainted truth effect, we find valid warnings help people reject misleading information, but we do not find that individuals are able to fully overcome the effect of misinformation and remember all of the correct information. Our findings generally align with the few studies that have previously examined this topic. However, our use of a diverse subject pool and political context reveals more muted effects and

insights into the influence of misinformation warnings on memory, memory uncertainty, and the perceived credibility of news that has been discounted by misinformation warnings.

Post-event Misinformation

Misinformation is broadly defined as "false or misleading information" (Lazer et al. 2018, p. 1094). Terms such as disinformation, fake or false news, and post-event misinformation refer to specific types of misinformation.[5] *Disinformation* is misinformation that is intentionally produced and spread to deceive people (Lazer et al. 2018; Wardle and Derakhshan 2017).[6] Often classified as a type of disinformation, *fake or false news* is fabricated information that assumes the guise of traditional news media but only in form, eschewing the organizational process or intent designed to produce accurate and credible information (Lazer et al. 2018; Wardle 2017, p. 1094). Finally, *post-event misinformation* is false information in the specific case where individuals have direct experience with an event but are later presented with misleading information about that original event. The post-event misinformation effect occurs when information inconsistent with an event and originating from another source enters an observer's recollection of that event (Szpitalak and Polczyk 2011, p. 140). While all types of misinformation are important to understand, our research focuses specifically on post-event misinformation in the context of political news to explore how retrospective warnings moderate post-event misinformation's effect on memory.

Historically, social cognition researchers have studied the post-event misinformation effect for the purpose of understanding eyewitness testimonies and criminal trials (e.g., Wyler and Oswald 2016). However, the post-event approach to misinformation can also be applied to political information and communication. While most of the political information received by the average individual is reprocessed through intermediaries (e.g., acquaintances, political elites, or media and journalistic sources), individuals often have existing knowledge of or experience with many of these reprocessed political events or issues. For example, people may watch a presidential debate and then read or watch commentary that summarizes and expands upon the debate.

Similarly, with the rise of video streaming and sharing on social media platforms, people can experience a political event almost directly and then later encounter the same event reprocessed through a post-event description, such as a news article. Moreover, the pluralistic nature of political communication often results in people

[5] With the politicization of the term "fake news," some scholars and organizations prefer to use the term "false news" (Lazer et al. 2018; Wardle and Derakhshan 2017; Tandoc Jr. et al. 2018). We use the terms interchangeably in this article.

[6] Tucker et al. (2018) define disinformation as encompassing an even wider range of information types found online including "fake news," rumors, factual information that is purposely misleading, inadvertently incorrect information, politically slanted information, and "hyperpartisan" news. We prefer a more precise terminology that separates purposive from inadvertent deception.

seeing multiple presentations of the same event, roughly mirroring the original event and post-event description paradigm. Whether the result of calculation or error, any reprocessing of information increases the likelihood that the information will be biased and misleading, thus opening individuals to the misinformation effect in the realm of political information.

Hundreds of studies over several decades have tackled the topic of the post-event misinformation effect (Ayers and Reder 1998; Blank and Launay 2014; Loftus 2005). In the 1970s, Elizabeth Loftus and colleagues were among the first to explore how eyewitness stories could be distorted by suggestive forensic interview practices (Loftus 1975; Loftus et al. 1978). Loftus et al. (1978) discovered that exposing people to misinformation about an event they had previously witnessed altered their ability to recognize details from the original event. This finding, referred to as the misinformation effect, was replicated in many studies under a wide range of conditions (for reviews see Ayers and Reder 1998; Chrobak and Zaragoza 2013; Loftus 2005; Frenda et al. 2011). Generally, a three-stage paradigm is used to investigate the misinformation effect. Participants are first shown an original event, then exposed to misleading information, and finally have their memory of the original event assessed, through either recognition or recall memory tests.

Misinformation Warnings and the Tainted Truth Effect

A subset of research on the misinformation effect explores whether the negative effects of misinformation on memory can be reversed, or at least minimized (e.g., Blank and Launay 2014; Chambers and Zaragoza 2001; Christiaansen and Ochalek 1983; Eakin et al. 2003; Echterhoff et al. 2005; Ecker et al. 2010; Wright 1993). For example, one of the earliest studies on the effects of misinformation warnings conducted by Dodd and Bradshaw (1980) found identifying the source of the misinformation as biased dramatically reduced the effect of misleading information on eyewitness memory. In the field of political science, a related body of literature also scrutinizes the causes, implications, and difficulty of countering *political* misinformation for topics, including the 2010 health care reform (Berinsky 2015; Nyhan 2010); climate change (van der Linden et al. 2017); campaign advertisements and political candidates (Amazeen et al. 2018; Cappella and Jamieson 1994; Pfau and Louden 1994; Thorson 2016; Wintersieck et al. 2018); political news (Clayton et al. 2019); and governmental policies, actions, and politically relevant data (Pennycook et al. 2018; Weeks 2015).[7] Under some conditions, warnings of misinformation can help individuals counter the effects of misinformation on attitudes and memory, but the corrections are often only partial, with long-lasting negative effects on trust (Cook and Lewandowsky 2011; Huang 2015; Lewandowsky et al. 2012; Nyhan and Reifler 2012). Warnings may even produce a boomerang or backfire effect and lead to misinformation becoming more deeply entrenched in memory when corrections

[7] See Flynn et al. (2017); Tucker et al. (2018) for a more comprehensive review of political misinformation research.

conflict with personal worldview or ideology (Nyhan and Reifler 2010). In a meta-analysis of 25 studies on retrospective warnings and post-event misinformation, Blank and Launay (2014) found retrospective warnings were only somewhat effective, on average reducing the post-event misinformation effect by half.

In addition to imperfectly counteracting misperceptions, misinformation warnings can produce other, often unintended, consequences. Although few in number, some studies outside of political science have investigated how misinformation warnings can extend beyond the intended target of misinformation and negatively influence surrounding information and memories. For example, Greene et al. (1982) discovered participants who were warned that post-event information came from an untrustworthy source were less likely to recognize events that were *correctly* described in the post-event description, compared to a no warning condition. Similarly, Meade and Roediger (2002) found warnings of an unreliable co-witness reduced recall of correct items reported by the co-witness.

Green et al. (1982) and Meade and Roediger (2002) noted the negative effects of warnings on memory, but these findings were not the primary focus of their research. Drawing on the research of Greene et al. (1982) and Meade and Roediger (2002), Echterhoff et al. (2007) deliberately began to study misinformation warnings' potentially adverse influence on correct memories, which they defined as the *tainted truth effect*. They found that when warned about misinformation, participants were less likely to recognize events that were accurately described in a post-event description, especially when the items were somewhat peripheral or difficult to remember.

In their investigation of the tainted truth effect, Echterhoff et al. (2007) considered various proposed mechanisms that could drive the misinformation and tainted truth effects.[8] Echterhoff et al. argued that under certain circumstances, misinformation warnings will reduce the ability to remember original events because warned individuals are more likely to monitor information from a source that has been discredited by a warning. Increased skepticism leads any information that is associated with the untrustworthy source to be tainted and rejected in retrospect, regardless of whether it is true or false. We also propose that retrospective warnings fundamentally alter how people reconstruct memory. In the absence of misinformation warnings, individuals should naturally rely more on post-event descriptions of an event as they are more recent and accessible (Wyler and Oswald 2016; Zaller 1992). However, when these post-event descriptions become tainted by misinformation warnings, individuals will feel more uncertainty and engage in a memory reconstruction process that discounts and rejects more recent data that comes from the post-event description, including both misinformation and accurate information.

[8] See Loftus (1975) for the initial memory impairment theory that theorized original event detail memory as being overwritten by misinformation. Other proposed mechanisms developed as subsequent research found the misinformation effect could be reduced through non-informative warnings (e.g., Blank and Launay 2014; Belli and Loftus 1996; Hell et al. 1988, Loftus 1991; Mazzoni and Vannucci 2007; McCloskey and Zaragoza 1985; Zaragoza et al. 2006). More recent proposed mechanisms model memory as a reconstruction process. When memory is assessed, a variety of construction strategies may be used, many of which are subject to different cognitive biases (Mazzoni and Vannucci 2007; Wyler and Oswald 2016).

Only a few studies on the tainted truth effect emerged after the initial formal consideration of the phenomenon by Echterhoff et al. (2007). In a series of related experiments, Szpitalak and Polczyk (2010, 2011, 2012) drew on Polish high school and university student subject pools to replicate and test the misinformation and the tainted truth effects in the contexts of a radio debate on education reform and a historical lecture on Christopher Columbus. Clayton et al. (2019) also recently identified the need for further research on the tainted truth effect in the area of political misinformation warnings. While the tainted truth effect was not the central hypothesis motivating their research, Clayton et al. (2019) found general warnings shown to participants before they read a set of headlines reduced the credibility of both truthful and untruthful headlines.

Our experiment contributes to the relatively understudied topic of the tainted truth effect by replicating and extending Szpitalak and Polczyk's (2011) study of misinformation, retrospective warnings of misinformation, and memory. Figure 1 illustrates a flow chart of the experimental design employed by Szpitalak and Polczyk (2011) to investigate the tainted truth effect. In Szpitalak and Polczyk's study, participants experienced an event (audio lecture on Christopher Columbus' expedition), read a description of the event following a lapse in time, and were tested on their memory of the original event.[9] Within this general design, participants were exposed to two main experimental manipulations: the first manipulation varied the content of the post-event description, and the second varied the presence of a retrospective warning of misinformation.

While all participants observed the exact same original event, the informational content of the written post-event description differed across three experimental description conditions. In the Control Condition, the post-event description was a vague summary of the original event with no review of the specific facts on which they were later tested. In the Information Condition, the post-event description provided an accurate review of precise facts seen in the original event that were also included in the final memory test. Finally, in the Misinformation Condition, the same set of detailed facts were presented to the participant in the post-event description, but a proportion of these facts were changed so they no longer accurately described the original event. The final manipulation altered whether a warning of misinformation followed the post-event description (Warning Condition and No Warning Condition).

Methods

Our study replicates the study design of Szpitalak and Polczyk (2011) but expands upon their research by examining the misinformation effect and post-warnings in the context of political news and testing the experiment through an online survey experiment with a more diverse subject pool.

[9] Although the results are applied to forensic science, the actual content examined was historical in nature.

Participants

Our online survey experiment was conducted from April 26 to 28, 2017 among adult U.S. participants recruited through Amazon Mechanical Turk (MTurk).[10] While the MTurk user population is not a representative sample of U.S. citizens, there is ample research suggesting it is a viable setting for survey experiments (Berinsky et al. 2012; Casler et al. 2013; Coppock 2018; Horton et al. 2011; Mullinix et al. 2015), and it is at least more diverse than the traditional experimental subject pool based on college students (Buhrmester et al. 2011).The sample of 434 participants used in our analyses is relatively diverse and comparable to the U.S. population (median age group is 35–54), although the sample we use is significantly more female (65% female) and more educated (52.1% have a bachelor's degree or greater).[11]

Procedure

To reduce respondent confirmation bias, the study was presented to participants under a cover story of "Color and Memory" (Podsakoff et al. 2012). Participants were told the purpose of the research was to "advance our understanding of the role of color in processing video material." Following brief instructions, participants were presented with the original event, a four minute CSPAN video recording of three U.S. House Representatives giving short speeches on the repeal of the Affordable Care Act, the UN resolution condemning Jewish settlement of the West Bank, and on the opening of the New York City subway.[12] These one-minute speeches

[10] In order to be eligible for the study, MTurk workers had to use a U.S. IP address, be over the age of 18, have a 95% or high approval rating for previous MTurk projects (HITs), and completed at least 50 other projects via Mturk. On April 26, 2017, eighty-three individuals participated in our study for a compensation of $.30 per subject. Realizing we underestimated the study's completion time, the compensation was raised to $.50 per subject the next two days while the study was open. Our substantive results remain when participation date/compensation amount is included as a control variable in the respective models.

[11] One difficulty of conducting experiments through online surveys is ensuring that participants actually receive the experimental treatments. Anticipating a degree of technical problems and insufficient exposure to experimental materials, our study measured both the technical experience of participants and time they spent on critical materials. A total of 549 participants entered the study, but 115 participants were dropped due to non-response or insufficient exposure to the main experimental treatments. Three participants entered the study but then exited immediately after reading the initial instructions, sixty-nine participants had technical problems viewing or did not view the entire video containing the original event materials, and forty-three people spent only ten seconds or less reading the post-event description. While these individuals cannot be included in the analysis, their failure to participate could introduce selection bias if they would have responded differently to the information and warning manipulations. The excluded participants did differ significantly from those who remained in the sample. Excluded participants were more likely to be younger males who had graduated from college but who had a lower need for cognition, read the news fewer days in a week, and knew less about politics. A more detailed analysis of the excluded participants and comparisons between age, education, and gender characteristics for the sample and census populations can be found in the supplementary information Tables SI-1–SI-3.

[12] The three January 4, 2017 one-minute U.S. House floor speeches for Representatives Bustos (D-Illinois, 17th District), Poe (R-Texas 2nd District) and Maloney (D-New York, 12th District) can be viewed here: https://www.c-span.org/video/standalone/?c4666248. For the full transcripts and more information on the original event materials, see the online supplementary information.

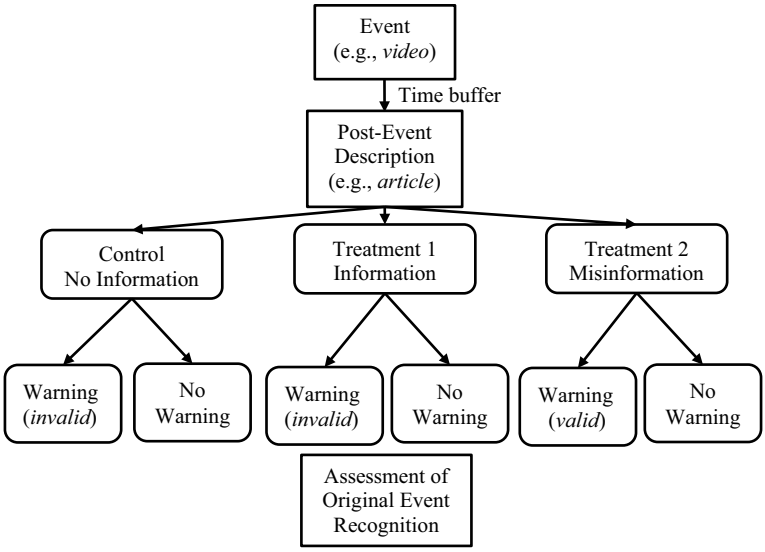

Fig. 1 Overview of Szpitalak and Polczyk's (2011) experimental approach

were selected because they covered a range of political issues (health care, foreign policy, and a regional infrastructure issue) presented by congressional members of both parties. To create a buffer period between the original event and post-event description, participants were asked to answer a set of 22 unrelated questions about their personal political positions and other basic demographic information after they viewed the video.

Participants were then randomly assigned to one of six conditions: a post-event description condition was crossed with exposure to a retrospective misinformation warning condition in a 3 × 2 between-subjects design (three description conditions: Control, Misinformation, Information by two warning conditions: No Warning, Misinformation Warning). Following the buffer period, participants were randomly exposed to one of three possible post-event descriptions (fabricated news articles) that had the same basic format but differed slightly in their content. In the Control Condition, the news article provided only a vague description of the original event/CSPAN video. In the Information Condition, specific facts from the floor speeches were inserted into the news article. In the Misinformation Condition, a subset of the specific facts was altered so the details no longer correctly reflected the original CSPAN video content. Each news article was formatted to look like a real article with a vague but plausible source: Jane Ross, a staff member the *Globe*. See the online supplementary materials for the entire news article transcript used in the description conditions.

Only a subset, rather than all facts, were manipulated in the Misinformation Condition to ensure the misinformation treatment was subtle and unlikely to lead people to reject the misinformation without any specific warning. After reading the news article, people were randomly assigned to one of two conditions. In the Warning

Condition, participants saw a misinformation warning, "warning: some of the information presented in the news article you read was inaccurate."[13] Participants in the No Warning Condition did not receive this warning. All survey questions and treatment materials are available in the online supplementary information.

After exposure to the post-event description and warning experimental materials, participants completed a recognition memory test of the 20 facts that were drawn from the CSPAN clip and described in the treatment conditions' news article. Eleven of the 20 factual questions corresponded to the 11 experimental facts that were altered to be misleading in the Misinformation Condition. The other 9 questions asked about the 9 fixed facts that were held constant in the news articles across all description conditions.

In the memory test, participants were asked to identify which one of four response options corresponded most closely to the information seen in the CSPAN video clip. Each question provided the accurate response option, two inaccurate options, and a "none of the answers are correct" option.[14] For the 11 experimental fact questions, one of the inaccurate options was the misinformation seen by participants in the Misinformation Condition. Following the memory test, participants were asked to rate the credibility of both the CSPAN video clip (original event) and the news article (post-event description) using an 11-item credibility measure. The full question and response wording and study material details can be found in the supplementary information.

Measures and Design-Specific Expectations

The primary dependent variable examined in this study is *memory score*: the ability to recognize information seen in the video (original event). Original event recognition memory scores are calculated as the percentage of test questions for which the participant correctly identified the response that corresponded to the original event information. Memory scores for both the 9 fixed fact subset and 11 experimental fact subset were calculated.

In alignment with the expectations of Szpitalak and Polczyk (2011), we anticipate that exposure to misinformation should lower memory score.[15] However, when respondents view accurate information in the news article, their ability to recognize the accurate information from the video should increase.[16] To investigate the effect of the news article on memory, we only consider respondents in

[13] The word "warning" was presented in red font to help draw the readers' attention.

[14] The presentation order of questions and response options was randomized.

[15] This expectation applies only to the experimental subset memory scores which correspond to the subset of details that were manipulated to be misleading in the Misinformation Condition.

[16] In the Information Condition, all news article details were accurate and the corresponding memory scores measured for both the fixed and experimental subset reflect exposure to accurate information. For respondents in the Misinformation Condition, the fixed fact subset news article details also correctly reflected the original video details. Therefore, the expectation that accurate information will improve memory score can also be considered for the fixed subset memory scores of individuals in the Misinformation Condition.

conditions without any misinformation warnings. Memory scores of respondents in the treatment conditions are compared to the memory scores of respondents in the Control Condition who only read vague post-event description news article. Formally, these expectations constitute the following two hypotheses as applied to our particular experimental design and measures:

Hypothesis 1a (Misinformation Effect) Exposure to misleading information in the post-event description news article is expected to reduce memory recognition of the original event video details: Memory scores (experimental fact subset) are expected to be lower in the Misinformation & No Warning Condition compared to the Control & No Warning Condition.

Hypothesis 1b (Information Effect) Exposure to accurate information in the post-event description news article is expected to increase the memory recognition of the original event video details: Memory scores (experimental and fixed fact subset) are expected to be higher in the Information & No Warning Condition compared to the Control & No Warning Condition. Memory scores (fixed fact subset) are expected to be higher in the Misinformation & No Warning Condition compared to the Control & No Warning Condition.

Assuming misinformation negatively affects memory score, we also expect warnings of misinformation will improve original event memory as warned individuals try and reject misleading information. First, memory scores should be higher for respondents who were exposed to misleading information in the news article and then later presented with a misinformation warning. Second, these more valid warnings of misinformation should also reduce the selection of the memory test response option that corresponds to the misleading information they were shown. The *misinformation score* is the percentage of experimental facts questions for which the participant selected the answer that corresponds with the misleading fact shown in the news article rather than the accurate information that was presented in the CSPAN video. If warnings make it easier to discard inaccurate information, respondents receiving valid warnings should have lower misinformation scores than individuals receiving the misinformation condition but no warning.

Hypothesis 2a (Warning and the Memory Performance) Exposure to a valid retrospective misinformation warning will increase the ability to correctly recognize original event details: Memory scores (experimental subset) are expected to be higher in the Misinformation & Warning Condition compared to the scores in the Misinformation & No Warning Condition.

Hypothesis 2b (Warning and the Misinformation Recognition) Exposure to a valid retrospective misinformation warning will reduce the incorrect recognition of misinformation as original event information: Misinformation scores (experimental subset) are expected to be lower in the Misinformation & Warning Condition compared to the scores in the Misinformation & No Warning Condition.

Because we expect misinformation warnings can contaminate accurate information, warnings should lead to the tainted truth effect even when they are invalid and no misinformation is present in the news article. When individuals are warned of misinformation, we anticipate worse memory scores as accurate information is rejected.

Given the design and fact subset structure of our study, the tainted truth effect hypothesis can be examined from multiple angles. Specifically, the tainted truth effect hypothesis logically leads us to test how misinformation warnings moderate all three components of the Information Effect considered in Hypothesis 1b. In the Information Condition, the post-event description is completely accurate, so the warning is invalid for both the fixed and experimental memory score subsets. In the Misinformation Condition, the warnings, while valid given the presence of misinformation, are still not completely valid due to their general, imprecise wording and potential for spillover. Therefore, the fixed facts (i.e., accurate information) have the potential to be tainted by the misinformation warning and rejected by respondents. A decrease in the memory score for the experimental fact subset (Information Condition) or the fixed fact subset (Information and Misinformation Conditions) will provide evidence that biased and inefficient warnings make it more difficult for respondents to recognize accurate information.

Hypothesis 3 (Tainted Truth Effect) Exposure to an invalid retrospective misinformation warning will reduce the ability to correctly recognize original event details: Memory scores (experimental and fixed subset) are expected to be lower in the Information & Warning Condition compared to the scores in the Information & No Warning Condition. Memory scores (fixed subset) are expected to be lower in the Misinformation & Warning Condition compared to the scores in the Misinformation & No Warning Condition.

While warnings aim to enable people to discard misinformation and correctly recognize original event material, warnings of misinformation may simply lead people to feel more uncertain about their memories. Specifically, people who are exposed to misinformation warnings may gravitate toward the response option, "none of the answers are correct," as they deal with more recognition confusion. An *uncertainty score* is calculated as the percent of the experimental and fixed fact subsets for which the participant selected the "none of the answers are correct" option. If warnings caused individuals to feel more confused and uncertain about their memory, we should see larger uncertainty scores in the Warning Conditions relative to the No Warning Conditions for all fact subsets and in all conditions.

Hypothesis 4a (Warning and Information Uncertainty) Exposure to a misinformation warning will increase memory uncertainty: Uncertainty scores (i.e., frequencies of selecting the "none of the answers are correct" response option) are expected to be higher in the Warning Conditions compared to the No Warning Conditions in all post-event description conditions.

Finally, participants were asked to evaluate how well eleven adjectives described the news article on a five-point Likert scale from "Not very well" to "Extremely well." A *credibility score* was calculated as the average of a participant's response to these eleven items (believable, accurate, trustworthy, biased [reverse coded], reliable, authoritative, honest, valuable, informative, professional, interesting). Even when the news article's source is not specifically mentioned in the misinformation warning, we expect participants to hold the source responsible for the veracity of the information. Warnings are expected to always reduce the perceived credibility of the news article (Hypothesis 4b).[17]

Hypothesis 4b (Perceived Credibility) Exposure to a misinformation warning will reduce the perceived credibility of the post-event description that is targeted by the warning: News article credibility scores are expected to be lower in the Warning Conditions compared to the No Warning Conditions in all post-event description conditions.

The tainted truth and warning effect expectations are critically tied to how misinformation warnings alter perceptions of the source. All information associated with a source connected to misinformation allegations become tainted, leading to the possibility that accurate information will be cast out with the false.

Results

Before formally testing each hypothesis, it is useful to consider the size of the treatment effects through summary statistics broken down by experimental condition for each relevant variable. Figure 2 shows the average memory scores within each description and warning condition for both the fixed and experimental subsets.[18] In the Control Conditions, the average participant is able to correctly recognize 59% of the original event items for both fixed and experimental question subsets shown in panels a and b. In panel b of Fig. 2, we see, on average, people who read misleading information in the news article only recognize 46% of the experimental subset's memory questions. This negative effect of misinformation on memory is also reflected in misinformation scores as shown in Fig. 3. On average, individuals who were exposed to misinformation but not warned about it incorrectly reported the misleading information as what they had seen in the original event video for 33% of the experimental subset questions (compared to 18% for people in the pure control condition). The valid warning of misinformation does seem to improve memory but only slightly, with the experimental subset mean memory score for people in the Warning & Misinformation Condition increasing to 52% (from 46% in the No

[17] The credibility of the original event CSPAN video was also measured and the index calculated. We do not expect the original event credibility to be significantly different over the description and warning conditions.

[18] Complete descriptive statistic for all measures are available in the online supplementary information.

Warning & Misinformation Condition) and the misinformation score decreasing to 25% (from 33%).

Conversely, exposure to accurate information in the news article boosts recognition memory. When people are exposed to accurate information in a news article (e.g., the memory scores for fixed experimental subsets in the Information Condition and fixed subset for the Misinformation Condition), average memory score jumps to around 72–76%. However, in these cases where the information in the news article was correct, subsequent warnings that there was misleading information in the news article suppress the memory scores down to 68–70%. This downward move in memory performance aligns with the expected direction of the tainted truth hypothesis, but the shift is marginal and the informed but warned memory scores are still higher than the 59% accuracy obtained in the Control Condition. At first glance, it looks like warnings of misinformation do not completely eradicate the benefits of accurate post-event information.

To formally test whether information, misinformation, and warnings of misinformation move memory in the expected directions, we interact warning and description condition indicators in three OLS regression models for the fixed and experimental memory score subset dependent variables. The first three models in Table 1 present the estimates used to test our three main hypotheses. In these models, participants in the Control & No Warning Condition serve as the baseline comparison group. Consistent with Hypothesis 1a, misinformation exposure reduces recognition accuracy as seen in the negative effect of misinformation on memory score in Model 1 ($\beta_{misinformation} = -12.98$, $se = 3.14$, $p < 0.001$) and positive effect on misinformation score in Model 3 ($\beta_{misinformation} = 14.61$, $se = 2.25$, $p < 0.001$). Hypothesis 1b, the prediction that accurate information in the news article increases original event recognition memory, is also supported by the positive and significant effect of information on memory score in Model 1 ($\beta_{information} = 15.94$, $se = 3.11$, $p < 0.001$) and Model 2 ($\beta_{information} = 16.94$, $se = 3.44$, $p < 0.01$).

The insignificant interaction terms in Models 1–3 of Table 1 reveal that the effects of warning on memory in the Information and Misinformation Conditions are not significantly different from the null effect found in the Control Condition.[19] However, while the warning effects in the Information and Misinformation Conditions are not statistically different from the effect established in the Control Condition, warning effects do emerge within the post-event description treatment conditions.

Columns 1–3 of Table 2 present the marginal effects of warning on memory and misinformation scores calculated from the estimates of Table 1. The marginal effects of warning on memory and misinformation scores are visually displayed in Fig. 4. The effects of valid warnings on memory performance (Hypothesis 2a) and misinformation endorsement (Hypothesis 2b) are visible in the green/triangular treatment

[19] Because there is no information to discard in the Control Condition where the news article offered only a vague description of the video, we did not expect the misinformation warning to alter memory scores in the Control Condition. However, it is possible that warning could heighten attention and thus improve the quality of memory reconstruction. This possibility that warnings affect memory through increased attentiveness does not hold in the data. Being warned about misinformation does not significantly alter memory performance (as seen in the insignificant $\beta warning$ in Models 1 and 2) or misinformation memory (Model 3) for people in the Control Condition.

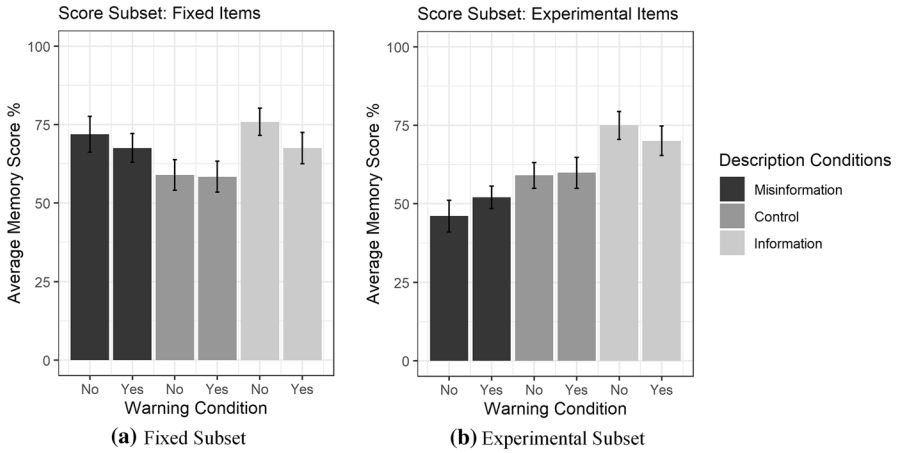

Fig. 2 Average memory score by condition

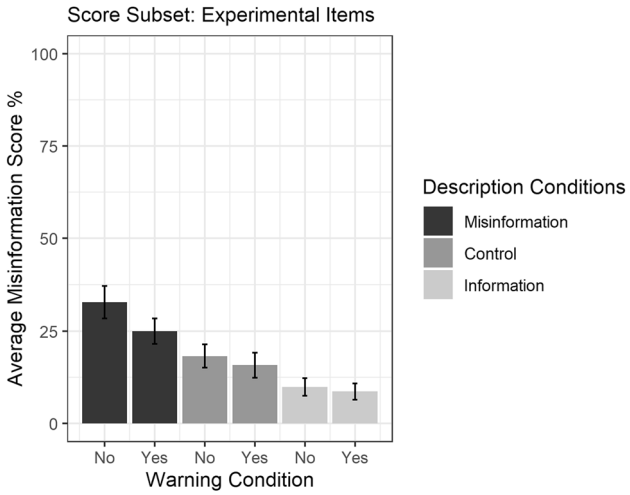

Fig. 3 Average misinformation score by condition, experimental subset

marginal estimates. We see in the right panel of Fig. 4, compared to participants in the misinformation condition who received no warning, those who were warned that they had been exposed to misleading information in the news article were significantly less likely to select the misleading information ($\beta_{warning} + \beta_{misinformation \times warning} = -7.90$, $se = 2.32$, $p < 0.01$).[20] However, rejection of misinformation does not fully

[20] While the insignificant interaction term in Model 3 of Table 1 ($\beta_{misinformation \times warning} = -5.43$, $se = 3.26$, $p = 0.10$) suggests this negative effect of warning on misinformation endorsement is not significantly different from the null effect of warning found in the Control Condition, the marginal effect of warning in the Misinformation Condition *is* significantly different from the null effect of warning found in the Information Condition ($\beta_{information \times warning} - \beta_{misinformation \times warning} = 6.68$, $se = 3.23$, $p = 0.04$).

Table 1 The effects of information and warning on memory score, misinformation, uncertainty score, and credibility scores

	Dependent variables						
	Memory score	Memory score	Misinfo. score	Uncertainty score	Uncertainty score	Credibility score	Credibility score
	Exp	*Fixed*	*Exp*	*Exp*	*Fixed*	*Video*	*Article*
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Warning	0.84	−0.57	−2.46	1.69	2.97	−0.01	−0.59**
	(3.19)	(3.52)	(2.29)	(2.28)	(2.42)	(0.14)	(0.15)
Misinformation	−12.98***	12.92***	14.61***	−1.13	−3.84	0.12	−0.23
	(3.14)	(3.47)	(2.25)	(2.24)	(2.38)	(0.14)	(0.15)
Information	15.94***	16.94**	−8.37***	−5.64*	−6.11**	0.01	0.2
	(3.11)	(3.44)	(2.23)	(2.23)	(2.36)	(0.14)	(0.14)
Misinformation × warning	5.16	−3.78	−5.43	0.65	−0.59	0.06	−0.03
	(4.54)	(5.01)	(3.26)	(3.25)	(3.44)	(0.20)	(0.21)
Information × warning	−5.75	−7.83	1.24	5.36+	4.5	0.08	0.21
	(4.47)	(4.94)	(3.21)	(3.20)	(3.39)	(0.20)	(0.21)
Constant	59.06***	58.98***	18.21***	13.46***	13.36***	3.61**	3.42**
	−2.15	−2.38	−1.54	(1.54)	(1.63)	(0.10)	(0.10)
Observations	434	434	434	434	434	434	433
R^2	0.21	0.08	0.28	0.03	0.04	0.01	0.14
Adj. R^2	0.20	0.07	0.27	0.02	0.03	−0.01	0.13

The baseline comparison group is the Control No Warning Condition. Estimates were obtained through OLS Regression. Memory, Misinformation, and Uncertainty Scores range from 0 to 100%. Credibility scores range from 1 to 5 points

*$p<0.05$; **$p<0.01$; ***$p<0.001$

Table 2 Marginal effects of warning by post-event description conditions

	Memory score *Exp* (1)	Memory score *Fixed* (2)	Misinfo. score *Exp* (3)	Uncertainty score *Exp* (4)	Uncertainty score *Fixed* (5)	Credibility score *Video* (6)	Credibility score *Article* (7)
Misinformation	6.00†	−4.35	−7.90**	2.34	2.38	0.05	−0.63***
	(3.23)	(3.57)	(2.32)	(2.31)	(2.45)	(0.14)	(0.15)
Control/no information	0.84	−0.57	−2.46	1.69	2.97	−0.01	−0.59***
	(3.19)	(3.52)	(2.29)	(2.28)	(2.42)	(0.14)	(0.15)
Information	−4.91	−8.39*	−1.22	7.05**	7.47**	0.08	−0.38**
	(3.13)	(3.46)	(2.25)	(2.24)	(2.38)	(0.14)	(0.15)

The baseline comparison groups are the no warning conditions within each respective description condition. Marginal effect estimates were obtained from the OLS Regression presented in Table 1. Memory, misinformation, and uncertainty scores range from 0 to 100%. Credibility scores range from 1 to 5 points

†$p < 0.10$; *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$

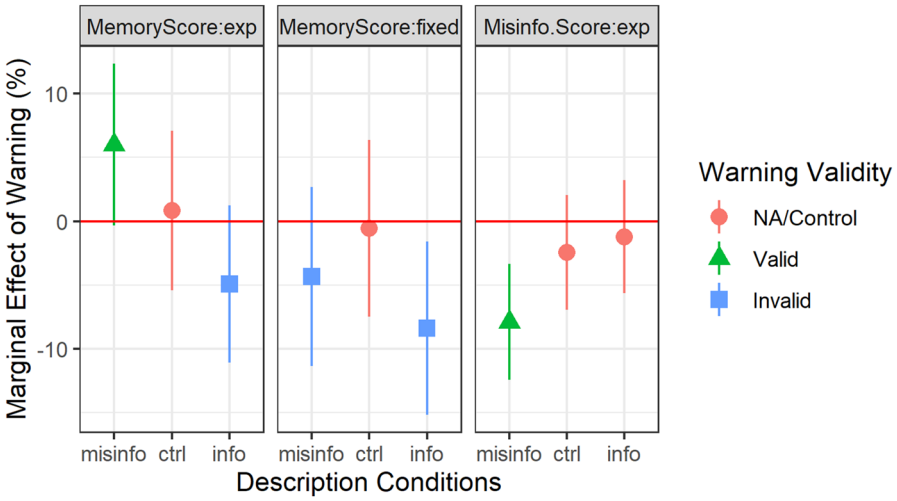

Fig. 4 Marginal effect of warning on memory by description condition

translate to correct identification of original event information. A complete correction would require a 13 point change to bring the memory score up to the level found in the control conditions (see Table SI-5 in the supplementary information for memory scores across the conditions). Our results find that while warned individuals reject the news article misinformation, they still struggle to remember the correct details they saw earlier in the video. On the left panel of Fig. 4, we see memory scores for individuals exposed to misleading information and then warned improve only 6 percentage points. This is consistent with Blank and Launay's (2014) finding that retrospective warnings usually only reduce the post-event misinformation effect by half. Even though the correction effect is not quite significant at the 0.05 level, it is positive and substantively large, indicating participants seek to counter misinformation when they have been alerted to its presence ($\beta_{warning} + \beta_{misinformation \times warning} = 6.00$, $se = 3.23$, $p = 0.06$).[21] These results suggest that misinformation may have a more persistent influence on memory as correction attempts often fall short.

The effects of invalid warnings and tests of the tainted truth effect (Hypothesis 3) are presented in the left and middle panels of Fig. 4 by the blue/square symbols that plot the marginal effects of warning on memory score. Of the three

[21] Just as we should be cautious about over-interpreting $0.04 < p < 0.05$, we should not over-interpret a $p = 0.06$ in light of the substantively large findings. Additionally, the insignificant interaction term in Model 1 of Table 1 ($\beta_{misinformation \times warning} = -5.16$, $se = 4.54$, $p = 0.26$) suggests this positive effect of warning on memory is not significantly different from the null effect of warning found in the Control Condition, but the positive marginal effect of warning in the Misinformation Condition *is* significantly different from the negative effect of warning found in the Information Condition ($\beta_{information \times warning} - \beta_{misinformation \times warning} = -10.91$, $se = 4.50$, $p = 0.01$). The treatment conditions' marginal effects of warning on memory as displayed in the left panel of Fig. 4 are not different from the Control, but they are different from each other.

possible tests of the tainted truth effect, a significant finding only emerges for the fixed fact subset for individuals in the Information Condition (middle panel; ($\beta_{warning} + \beta_{misinformation \times warning} = -8.39$, $se = 3.46$, $p = 0.02$). For these same individuals in the Information Condition, warnings still suppress memory score for the experimental fact subset, but the effect size is not large enough to reach statistical significance ($\beta_{warning} + \beta_{misinformation \times warning} = -4.91$, $se = 3.13$, $p = 0.12$).[22] Similarly, participants in the Misinformation Condition are more likely to reject the valid information (fixed fact subset) when they are warned of misinformation, but the size of rejection is too small to be significantly different from those people in the No Warning & Misinformation Condition ($\beta_{warning} + \beta_{misinformation \times warning} = -4.35$, $se = 3.57$, $p = 0.22$).[23]

The final aspect of memory responses that our experimental design and data allows us to examine is uncertainty. Figure 5 presents the average uncertainty scores over the six conditions. Contrary to our expectations in Hypothesis 4a, warnings of misinformation do not appear to consistently alter uncertainty. Uncertainty increases slightly in all warning conditions, but warnings only significantly alter uncertainty in the Information Condition. When the information in the news article is completely correct as is found in the Information Condition, participants appear more confident in their memory compared to those who are informed but then exposed to an invalid warning. Informed and not warned individuals select the "none of the answers are correct" option for only around 7% of the questions, but when warned this number rises to 15%.

Models 4 and 5 in Table 1 estimate the effect of information, misinformation, and warnings on uncertainty scores. The corresponding marginal effects of warnings are presented in columns 4 and 5 in Table 2 and graphically displayed Fig. 6. In both models 4 and 5 of Table 1, while the level of uncertainty in the No Warning & Information Condition is significantly different from the uncertainty in the No Warning & Control Condition for both experimental ($\beta_{information} = -5.64$, $se = 2, 23$, $p = 0.012$) and fixed ($\beta_{information} = -6.11$, $se = 2.36$, $p = 0.01$) fact subsets, the upward shift in uncertainty produced by warnings is not significantly different from that found in the Control Condition (exp: $\beta_{information \times warning} = 5.35$, $se = 3.20$, $p = 0.094$;

[22] The insignificant interaction terms in Model 2 of Table 1 ($\beta_{misinformation \times warning} = -3.77$, se = 5.01, p = 0.45) and ($\beta_{information \times warning} = -7.83$, se = 4.94, p = 0.11) suggest this negative effect of warning on memory (fixed subset) is not significantly different from the null effect of warning found in the Control Condition. If the warnings lead to accurate information being rejected, we would expect to see a negative effect of warning on fixed subset memory in both of the Description treatment conditions. As expected, the negative effect of warning is not significantly different between the Misinformation and Information Conditions ($\beta_{information \times warning} - \beta_{misinformation \times warning} = -4.05$, $se = 4.97$, $p = 0.42$).

[23] While these multiple tests allow us to consider the tainted truth effect in different aspects of the design, as noted by Gelman and Stern (2006), these tests do not identify whether the differences between the tests are significant. Even though only one test reached statistical significance, their collective alignment in direction and substance builds a stronger case for the tainted truth effect. Also, the tainted truth effect remains statistically significant in a comparison of warned and not warned respondents in the Information Condition when fixed and experimental subsets are combined to create an overall memory score.

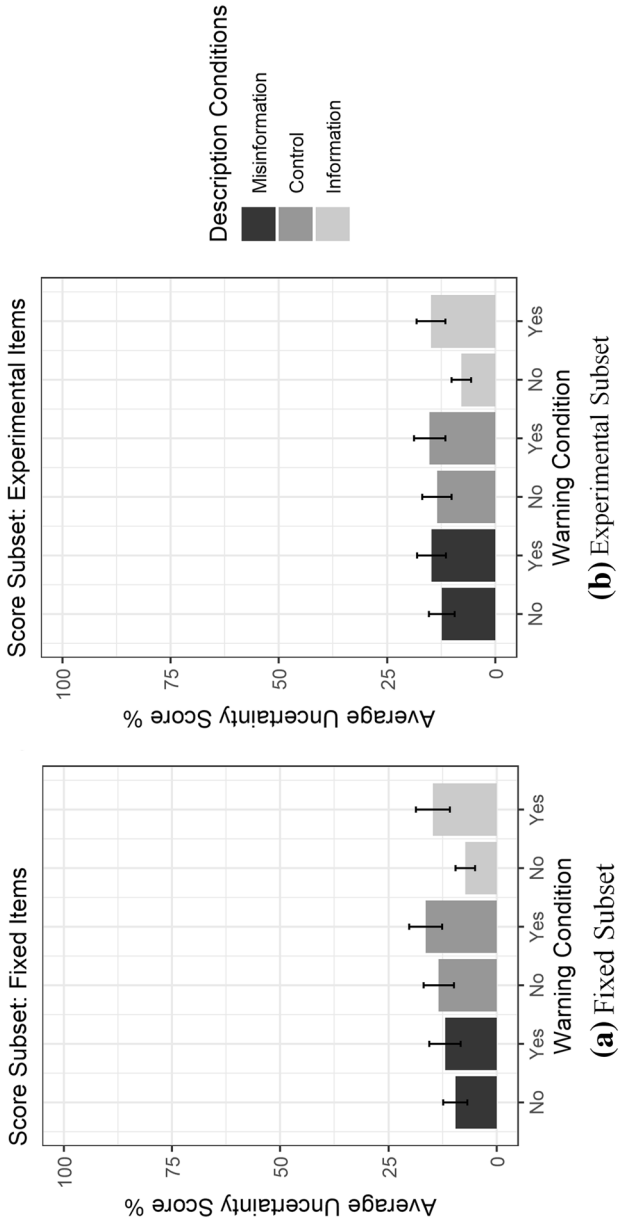

Fig. 5 Average memory uncertainty score by condition

fixed: $\beta_{information \times warning} = 4.50$, $se = 3.39$, $p = 0.185$).[24] However, as shown by the marginal effect estimates in Fig. 6, *within* the Information Condition, warnings significantly move the uncertainty score around 7 percentage points. In the absence of warning, individuals presented with an accurate news article in the Information Condition were more likely to be certain about their memory compared to those in the Control or Misinformation Conditions. But once exposed to a misinformation warning, individuals doubt their memory and response uncertainty jumps to average levels seen in the other condition.

Having considered all possible aspects of memory as influenced by information, misinformation, and misinformation warnings, we now turn to the primary mechanism proposed by Echterhoff et al. (2007): source monitoring. The rejection of misinformation, rejection of accurate information, and increase in memory uncertainty occur as general warnings taint all information, good and bad, that people associate with the allegedly misleading source. And, as seen in panel b in Fig. 7, average levels of the news article's credibility clearly decrease under warning conditions for all information conditions. For example, in the condition where the news article should have the most credibility (No Warning & Information Condition), we see the article has the same average credibility score as the CSPAN video displayed in panel b. Furthermore, as expected since the video content was held constant across all conditions, the credibility of the video does not significantly change across the conditions (See Model 6 of Table 1 and marginal effects in column 6 of Table 2).

In contrast to their perceptions of the video, respondents' perceptions of the news article credibility do significantly respond to the experimental treatments. Looking at the estimated effect of *warning* in Model 7 of Table 1, we see that a misinformation warning leads individuals in the Control Condition to view the news article as 0.59 points less credible (out of five points). The insignificant warning and description conditions interactions reveals that the significant negative effect of the misinformation warning on article credibility found in the Control Condition also occurs in the Information and Misinformation Conditions.

It is important to note that while warnings did reduce the credibility of the news article in all conditions, the news article credibility is not completely identical across all the baseline (No Warning) description conditions. The accurate news article in the No Warning & Information Condition is significantly more credible than the misleading article in the No Warning & Misinformation Condition ($\beta_{information} - \beta_{misinformation} = 0.43$, $se = 0.15$, $p = 0.004$). Participants were probably somewhat aware of the misinformation even without being exposed to a retrospective warning. While our design sought to keep the misleading information subtle by changing only a subset (the experimental facts) to be false, the lower levels of credibility in the unwarned Misinformation Conditions suggests the manipulation might have not been subtle enough. Even though overall credibility is significantly

[24] The positive marginal effect of warning on uncertainty in the Information Condition is also not significantly different from the marginal effect in the Misinformation Condition; exp: $\beta_{information \times warning} - \beta_{imisnformation \times warning} = 4.71$, $se = 3.22$, $p = 0.14$; fixed: $\beta_{information \times warning} - \beta_{imisnformation \times warning} = 5.09$, $se = 3.41$, $p = 0.14$.

higher in the Information Condition compared to the Misinformation Condition, within each description condition the warnings still produced significant drops, as seen in the statistically significant marginal effects of warning presented in column 7 of Table 2 and in the right panel of Fig. 8.[25]

Discussion

Our research replicates the relatively unexplored tainted truth effect and provides useful insights into how efforts to prevent misinformation can have unintended and negative consequences for memory. We find that invalid misinformation warnings can damage source credibility and cause people to reject accurate information that is associated with the tainted source. Warnings of misinformation can also cause people to feel more uncertain about their memory, especially when they were in fact not exposed to any information and the warnings are completely invalid. While valid warnings of misinformation enable people to reject false information, misdirected and imprecise warnings may counter the positive influence of misinformation warnings on memory.

In addition to extending the tainted truth effect to the domain of political communication, our research also provides an interesting launching point for exploring the complexity of invalid warnings of political misinformation. In a February 19, 2019 Twitter post, President Donald Trump alleged the existence of invalid misinformation warnings by saying: "The Washington Post is a Fact Checker only for Democrats. For Republicans, and for your all time favorite President, it is a Fake Fact Checker!" Although this reasoning may feel as though we are being pulled down the rabbit hole with Alice, it does broach several interesting questions: can misinformation warnings be countered and how does political and ideological congruence with the sources moderate these attempts? A recent public opinion study identified Republicans as more likely than Democrats to say fact-checking efforts by news organizations favor one side (Walker and Gottfried 2019). If the source of a misinformation warning is perceived as less credible, does it alter the effect of warnings and potential tainted truth effects? Our study begins to address the varied potential effects of misinformation warnings and we suggest this is a topic of inquiry ripe for exploration.

One clear practical implication for political psychology that stems from our tainted truth effect research is the recognition that misinformation warnings may have a dark side as they can lead people to feel more uncertain about, less trusting of, and more likely to reject accurate information. If invalid misinformation warnings have the potential to impede political knowledge, we need to more clearly identify what constitutes an invalid warning and when spillover effects are

[25] The negative marginal effect of warning on news credibility does not differ significantly between any of the Description Conditions as is seen in the insignificant interaction terms in Model 7 of Table 1 and the insignificant linear combination test that compares the interaction terms: $\beta_{informationXwarning} - \beta_{imisnformationXwarning} = 0.25$, se = 0.21, p = 0.24,

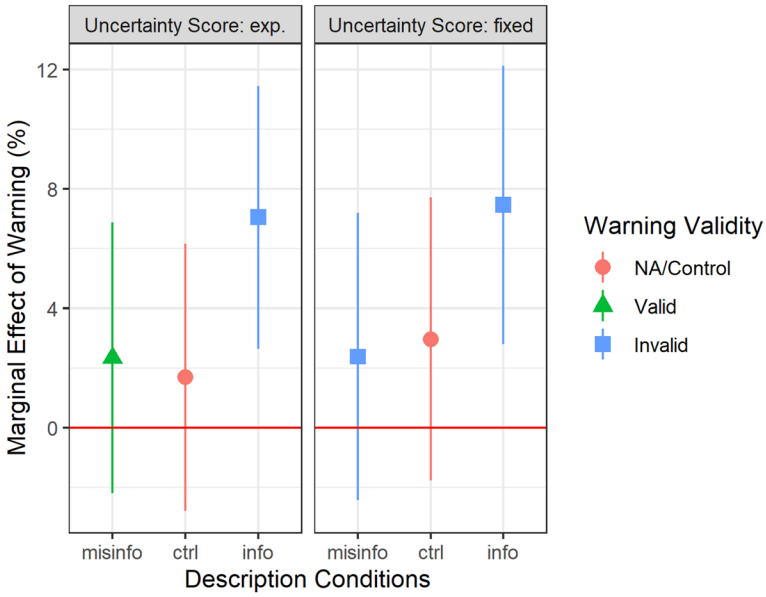

Fig. 6 Marginal effect of warning for memory uncertainty by description condition

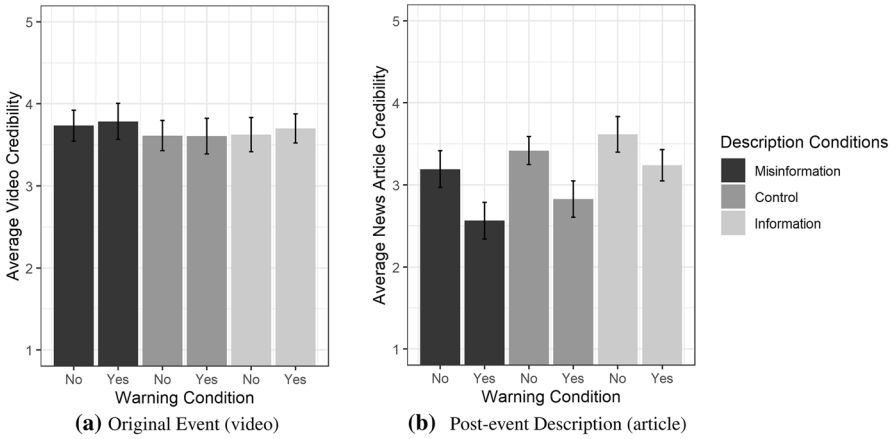

Fig. 7 Average credibility of original event (video) and post-event description (article) by condition

likely to occur. The quality of misinformation warnings needs to become a part of the dialogue surrounding investigation of misinformation in the realm of politics. Just as there has been an explosion of fact-checking organizations in the past decade as misinformation has become more salient, there may be a demand for comparable efforts that enhance the integrity of valid fact-checkers. In line with this is the need for further research on trust of fact-checking organizations and

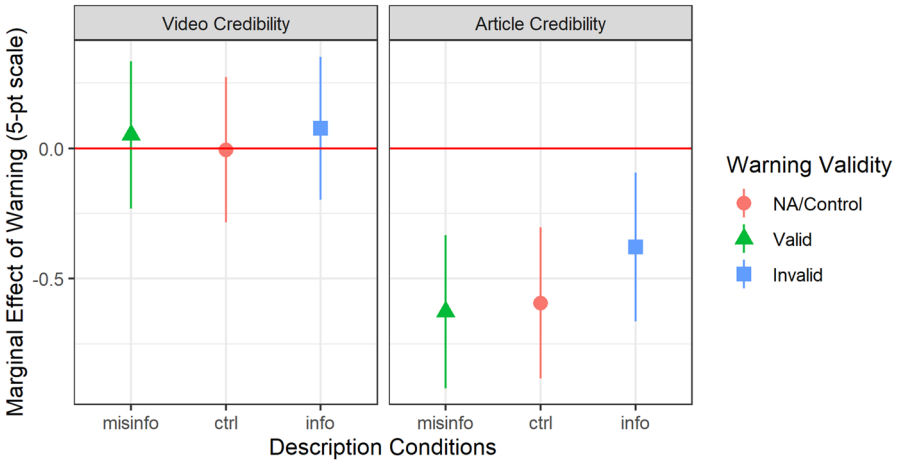

Fig. 8 Marginal effects of warning on source credibility by description conditions

other sources of misinformation warnings to better understand when misinformation may be more or less effective.

Several design choices made in our study could also be revisited in subsequent research. First, the original event and post-event description form and content may influence how information is processed and whether the tainted truth effect is amplified or minimized. In the study conducted by Szpitalak and Polczyk (2011), participants experienced the original event information in audio form while the post-event description was read. Our study presented a video original event and written post-event description. In the field of social cognition, misinformation has been introduced through various forms including direct personal interaction, written, and, sometimes, audio. When misinformation encounters are classified as direct (e.g., face-to-face, co-witness, social) versus indirect (e.g., written reports, non-social), Blank et al. (2013) found no clear difference in misinformation retention in the area of eyewitness reports. However, these findings may not hold in the area of warning effects and political communication. For example, warnings of misinformation may be less likely to taint good information if the post-event description comes in the form of a written news article compared to a radio or television news program if information is encoded more strongly through reading versus listening or watching.

Second, careful considerations of the type of (mis)information accepted or rejected in the face of retrospective warnings should be addressed in future studies of the tainted truth effect. While our design sought to incorporate a wide range of political topics including health care, foreign policy, and distributive politics, the facts used in the memory test were mostly novel and moderately peripheral. We chose to test recognition memory of these details for several reasons. First, details were relatively obscure (e.g., how many jobs a new subway generated), thus minimizing the likelihood participants would have prior exposure and heterogeneous ability to remember them (Pennycook et al. 2018). Second, we chose to examine memory of moderately peripheral information to

minimize the likelihood of participants independently identifying our misinformation manipulations in the post-event description. However, the lower source credibility in the No Warning Misinformation Condition compared to the No Warning Information Condition leads us to doubt whether our details and misinformation manipulations were peripheral enough.

Furthermore, how easy or difficult it to remember information can alter the power of a warning. For very difficult, or peripheral, misinformation details where no memory exits, individuals engage in "best-guess" strategies in recognition memory tests and warnings make no difference (Wyler and Oswald 2016). On the other hand, when the information is very easy to remember, retrospective warnings may also have little influence as individuals are able to identify and correct for the misinformation at the time of exposure (Putnam et al. 2017). The misinformation effect and impact of subsequent warnings tend to be largest for moderately peripheral information. For details that are somewhat difficult to remember, misinformation is often undetected and recognition tasks are more prone to recency or familiarity bias which warnings can later mitigate (Wyler and Oswald 2016).

Our findings may have been muted because we chose to examine information that was too memorable or too peripheral. Table SI-4 in the online supplementary information suggests most items were only moderately difficult and the difficulty for the experimental and fixed subsets similar, but further research could specifically examine the effect of information difficulty and the tainted truth effect. Future extensions of our research could also examine how ideological congruence with either the original event information, source of the post-event description, or source of the retrospective warning alters the tainted truth effect. While the warnings in our study came from the researcher (warning source was not clearly specified), it would be interesting to see if motivated reasoning alters the tainted truth effect if Trump or some other source presents the misinformation warnings.

While we were able to test this effect on a sample that was reasonably diverse, our results may be limited by the substantial number of participants we had to exclude due to insufficient exposure to our experimental manipulations. Future research using a nationally representative sample and an experimental design that reduces attrition may reveal different effect sizes. Finally, our design examined recognition memory within a relatively short experiment. Participants took 16 min on average to complete the study. While the inclusion of a set of unrelated questions after the original event provided some buffer period between the original event and post-event description, a design that considers the tainted truth effect over longer time intervals between the original event, post-event description, warning, and memory test could shed greater understanding of the cognitive processes underlying the tainted truth effect. Altering the format of the memory test could also help provide more understanding of foundational mechanisms. For example, changing the memory test to forced choice and adding an additional question that measures memory uncertainty could identify the whether or not correction attempts are produced by confusion or enlightenment.

Conclusion

One of the basic assumptions of a well-functioning democracy is the presence of an educated and well-informed citizenry (Lewandowsky et al. 2012). At face value, misinformation threatens democratic proceedings if it can influence and shape public opinion and social decisions. Consequently, numerous studies and efforts have emerged to identify and counteract the effects of misinformation in journalistic settings and broader areas of political communication. Our research takes a step back from this fundamental problem to consider whether the efforts to combat misinformation in themselves may have negative side effects.

Our research replicates the tainted truth effect and extends it to the area of political news. Our findings cast much needed light on this phenomenon that has gathered only a little attention in the field of social cognition and even less in the area of political news and communication. Drawing on a relatively diverse sample, we reproduce the general results of prior studies of misinformation and warnings. We find clear evidence that post-event descriptions of prior events shape memory. When original events are twisted by misinformation in a subsequent news article, people are more likely to recognize the false information as the original event data and less likely to identify the correct facts. Conversely, exposure to a news article that provides an accurate retelling of an event experienced earlier boosts individuals' abilities to correctly remember original event items. When these news articles are then followed by statements warning individuals that the news articles contained some misleading information, we find several interesting developments in recognition memory. Although people try to correct for the misinformation, these efforts are often inadequate. Valid warnings lead people to try and discard the false data seen in the news article, but they still struggle to correctly remember the original event details.

Warnings of misinformation potentially hold other negative consequences for an informed citizenry. When the allegations of misinformation in the news article are invalid, people reject the accurate information, leading to the tainted truth effect. False warnings of misinformation reduce the credibility of legitimate news, decrease acceptance of useful news data, increase memory uncertainty, and impede original event memory. However, these negative effects of misinformation warnings on memory are constrained as the decrease is substantively small. We find the tainted truth effect does not completely erode the positive benefits of factual news on memory.

Our research finds that both valid and invalid retrospective warnings reduce news credibility and alter how news information is processed. Given the potential for misinformation warnings to impede the credibility and acceptance of real news, more attention and research on the tainted truth effect and other unforeseen negative consequences of general warnings of misinformation is needed. We join (Clayton et al. 2019) and others' recommendations that fact-checkers, news media, and political elites tread carefully when deploying general allegations and warnings of fake news and misinformation. While misinformation warnings are critical in combatting the negative effects of misinformation, it is important to

be cognizant of the many possible spillover effects from general warnings which may unintentionally damage real news institutions that support critical democratic processes.

Acknowledgements Other co-authors are or were undergraduate students at Carleton College (Yoichiro Ashida, Mitch Bermel, Sharaka Berry, Jeremy Brog, Ursula Clausing, Eveline Dowling, Maximilian Esslinger, M. Forsyth, Malcom Fox, Shayna Gleason, Lea Gould, Boluwatife Johnson, Schuyler, Kapnick, Mark Leedy, Calypso Leonard, Malekai Mischke, Isabel Storey, Oliver Wolyniec, Sol Yanuck), Spring 2017 POSC 226 course. We thank Carleton College, Department of Political Science and Dean of the College for generous funding support and the Headley Travel Fund Grant. We also thank the Department of Political Science at Brigham Young University-Provo for workshop travel support. We are grateful for the advice of Brendan Nyhan, Jessica Preece, Kent Freeze and two anonymous reviewers.

Data Availability Replication materials can be found on Harvard Dataverse at: https://doi.org/10.7910/DVN/TRR0DK

References

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives, 31*(2), 211–236.

Allcott, H., Gentzkow, M., Yu, C. (2019). *Trends in the diffusion of misinformation on social media. Technical report*. National Bureau of Economic Research. Retrieved April 23, 2019, from https://www.nber.org/papers/w25500.pdf.

Amazeen, M. A., Thorson, E., Muddiman, A., & Graves, L. (2018). Correcting political and consumer misperceptions: The effectiveness and effects of rating scale versus contextual correction formats. *Journalism & Mass Communication Quarterly, 95*(1), 28–48.

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

Belli, R. F., & Loftus, E. F. (1996). The pliability of autobiographical memory: Misinformation and the false memory problem. In D. C. Rubin (Ed.), *Remembering our past: Studies in autobiographical memory* (pp. 157–179). New York: Cambridge University Press.

Berinsky, A. J. (2015). Rumors and health care reform: Experiments in political misinformation. *British Journal of Political Science, 47*(2), 241–262.

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Political Analysis, 20*, 351–368.

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*(2), 77–88.

Blank, H., Ost, J., Davies, J., Jones, G., Lambert, K., & Salmon, K. (2013). Comparing the influence of directly vs indirectly encountered post-event misinformation on eyewitness remembering. *Acta Psychologica, 144*(3), 635–641.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science, 6*, 3–5.

Cappella, J. N., & Jamieson, K. H. (1994). Broadcast adwatch effects: A field experiment. *Communication Research, 21*(3), 342–365.

Casler, K., Bickel, L., & Hackett, E. (2013). Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. *Computers in Human Behavior, 29*, 2156–2160.

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

Christiaansen, R. E., & Ochalek, K. (1983). Editing misleading information from memory: Evidence for the coexistence of original and postevent information. *Memory & Cognition, 11*(5), 467–475.

Chrobak, Q. M., & Zaragoza, M. S. (2013). The misinformation effect: Past research and recent advances. In A. M. Ridley, F. Gabbert, & D. J. Rooy (Eds.), *Suggestibility in legal contexts: Psychological research and forensic implications* (pp. 21–44). West Sussex, UK: Wiley-Blackwell.

Clayton, K., Blair, S., Busam, J. A., Forstner, S., Glance, J., Green, G., et al. (2019). Real solutions for fake news? Measuring the effectiveness of generalwarnings and fact-check tags in reducing belief in false stories on social media. Political Behavior. doi: 10.1007/s11109-019-09533-0.

Cook, J., & Lewandowsky, S. (2011). *The debunking handbook*. St. Lucia: University of Queensland.

Coppock, A. (2018). Generalizing from survey experiments conducted on Mechanical Turk: A replication approach. *Political Science Research and Methods, 7*(3), 1–16.

Dodd, D. H., & Bradshaw, J. M. (1980). Leading questions and memory: Pragmatic constraints. *Journal of Verbal Learning and Verbal Behavior, 19*(6), 695–704.

Eakin, D. K., Schreiber, T. A., & Sergent-Marshall, S. (2003). Misinformation effects in eyewitness memory: The presence and absence of memory impairment as a function of warning and misinformation accessibility. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*(5), 813.

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

Echterhoff, G., Groll, S., & Hirst, W. (2007). Tainted truth: Overcorrection for misinformation influence on eyewitness memory. *Social Cognition, 25*(3), 367–409.

Ecker, U. K. H., Lewandowsky, S., & Tang, D. T. W. (2010). Explicit warnings reduce but do not eliminate the continued influence of misinformation. *Memory & Cognition, 38*(8), 1087–1100.

Flynn, D. J., Nyhan, B., & Reifler, J. (2017). The nature and origins of misperceptions: Understanding false and unsupported beliefs about politics. *Political Psychology, Supplement: Advances in Political Psychology, 38*(S1), 127–150.

Frankovic, K. (2016). Belief in conspiracies largely depends on political identity. *YouGov*. Retrieved April 22, 2019, from https://today.yougov.com/topics/politics/articles-reports/2016/12/27/belief-conspiracies-largely-depends-political-iden.

Frenda, S. J., Nichols, R. M., & Loftus, E. F. (2011). Current issues and advances in misinformation research. *Current Directions in Psychological Science, 20*(1), 20–23.

Funke, D. (2019). Facebook announces sweeping changes to its anti-misinformation policies. *Poynter*, April 10. Retrieved April 22, 2019, from https://www.poynter.org/fact-checking/2019/facebook-announces-sweeping-changes-to-its-anti-misinformation-policies/.

Gelman, A., & Stern, H. (2006). The difference between "significant" and "not significant" is not itself statistically significant. *The American Statistician, 60*(4), 328–331.

Graves, L. (2016). *Deciding what's true: The rise of political fact-checking in American journalism*. New York: Columbia University Press.

Graves, L., Nyhan, B., & Reifler, J. (2016). Field experiment examining motivations for fact-checking. *Journal of Communication, 66*(1), 102–138.

Greene, E., Flynn, M. S., & Loftus, E. F. (1982). Inducing resistance to misleading information. *Journal of Verbal Learning and Verbal Behavior, 21*(2), 207–219.

Grynbaum, M. M. (2019a). Buzzfeed news faces scrutiny after Mueller denies a dramatic Trump report. *The New York Times*, January 19. Retrieved June 8, 2019, from https://www.nytimes.com/2019/01/19/business/media/buzzfeed-news-trump-michael-cohen-mueller.html.

Grynbaum, M. M. (2019b). Trump discusses claims of 'fake news,' and their impact with New York Times publisher. *The New York Times*, February 1. Retrieved April 22, 2019, from https://nyti.ms/2DMIXwq.

Guess, A., Lyons, B., Montgomery, J. M., Nyhan, B., & Reifler, J. (2018a). Fake news, Facebook ads, and misperceptions: Assessing information quality in the 2018 U.S. midterm election campaign. Retrieved April 22, 2019, from https://www-personal.umich.edu/ bnyhan/fake-news-2018.pdf.

Guess, A., Nyhan, B., & Reifler, J. (2018b). Selective exposure to misinformation: Evidence from the consumption of fake news during the 2016 us presidential campaign. *European Research Council*, January 9. Retrieved April 22, 2019, from https://www.dartmouth.edu/~nyhan/fake-news-2016.pdf.

Hell, W., Gigerenzer, G., Gauggel, S., Mall, M., & Müller, M. (1988). Hindsight bias: An interaction of automatic and motivational factors? *Memory & Cognition, 16*(6), 533–538.

Horton, J. J., Rand, D. G., & Zeckhauser, R. J. (2011). The online laboratory: Conducting experiments in a real labor market. *Experimental Economics, 14*(3), 399–425.

Huang, H. (2015). A war of (mis)information: The political effects of rumors and rumor rebuttals in an authoritarian country. *British Journal of Political Science, 47*(2), 283–311.

Kang, C. (2019) Facebook's Hands-Off Approach to Political Speech Gets Impeachment Test. *The New York Times*, October 8. Retrieved January 3, 2019, from https://www.nytimes.com/2019/10/08/technology/facebook-trump-biden-ad.html.

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., et al. (2018). The science of fake news: Addressing fake news requires a multidisciplinary effort. *Science, 369*(6380), 1094–1096.

Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest, 13*(3), 106–131.

Loftus, E. F. (1975). Leading questions and the eyewitness report. *Cognitive Psychology, 7*(4), 560–572.

Loftus, E. F. (1991). Made in memory: Distortions in recollection after misleading information. In G. H. Bower (Ed.), *Psychology of learning and motivation* (Vol. 27, pp. 187–215). Cambridge: Academic Press.

Loftus, E. F. (2005). Planting misinformation in the human mind: A 30-year investigation of the malleability of memory. *Learning & Memory, 12*(4), 361–366.

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

Lyons, T. (2018a). Hard questions: What's Facebook's strategy for stopping false news? *Facebook Newsroom*, May 23. Retrieved April 22, 2019, from https://newsroom.fb.com/news/2018/05/hard-questions-false-news/.

Lyons, T. (2018b). Increasing our efforts to fight false news. *Facebook Newsroom*, June 21. Retrieved June 23, 2019, from https://newsroom.fb.com/news/2018/06/increasing-our-efforts-to-fight-false-news/.

Mazzoni, G., & Vannucci, M. (2007). Hindsight bias, the misinformation effect, and false autobiographical memories. *Social Cognition, 25*(1), 203–220.

McCloskey, M., & Zaragoza, M. (1985). Misleading postevent information and memory for events: Arguments and evidence against memory impairment hypothesis. *Journal of Experimental Psychology: General, 114*, 1–16.

Meade, M. L., & Roediger, H. L. (2002). Explorations in the social contagion of memory. *Memory & cognition, 30*(7), 995–1009.

Mosseri, A. (2017). Working to stop misinformation and false news. *Facebook for Media*, April 7. Retrieved April 22, 2019, from https://newsroom.fb.com/news/2017/04/working-to-stop-misinformation-and-false-news/.

Mullinix, K. J., Leeper, T. J., Druckman, J. N., & Freese, J. (2015). The generalizability of survey experiments. *Journal of Experimental Political Science, 2*(2), 109–138.

Nyhan, B. (2010). Why the "death panel" myth wouldn't die: Misinformation in the health care reform debate. *The Forum, 8*(1), 1–24.

Nyhan, B. (2019). Why fears of fake news are overhyped, February 22. *Medium*. Retrieved April 22, 2019, from https://medium.com/s/reasonable-doubt/why-fears-of-fake-news-are-overhyped-2ed9ca0a52c9.

Nyhan, B., & Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior, 32*(2), 303–330.

Nyhan, B., & Reifler, J. (2012). Misinformation and fact-checking: Research findings from social science. *Media Policy Initiative, New America Foundation*, February 28. Retrieved April 22, 2019, from https://www.newamerica.org/oti/policy-papers/misinformation-and-fact-checking/.

Pennycook, G., & Rand, D. G. (2017). The implied truth effect: Attaching warnings to a subset of fake news stories increases perceived accuracy of stories without warnings. Retrieved May 8, 2017, from https://tinyurl.com/y25rxlmc.

Pennycook, G., Cannon, T. D., & Rand, D. G. (2018). Prior exposure increases perceived accuracy of fake news. *Journal of experimental psychology, 147*(12), 1865–1880.

Pfau, M., & Louden, A. (1994). Effectiveness of adwatch formats in deflecting political attack ads. *Communication Research, 21*(3), 325–341.

Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2012). Sources of method bias in social science research and recommendations on how to control it. *Annual Review of Psychology, 63*, 539–569.

Putnam, A. L., Sungkhasettee, V. W., & Roediger, H. L., III. (2017). When misinformation improves memory: The effects of recollecting change. *Psychological Science, 28*(1), 36–46.

Silverman, C. (2016). This analysis shows how fake election news stories outperformed real news on facebook. *Buzzfeed News*, November 16. Retrieved April 19, 2019, from https://www.buzzfeednews.com/article/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook.

Silverman, C., & Singer-Vine, J. (2016). Most Americans who see fake news believe it, new survey says. *Buzzfeed News*, December 6. Retrieved April 22, 2019, from https://www.buzzfeednews.com/article/craigsilverman/fake-news-survey.

Silverman, C., Strapagiel, L., Shaban, H., Hall, E., & Singer-Vine, J. (2016). Hyperpartisan Facebook pages are publishing false and misleading information at an alarming rate. *Buzzfeed News*, October 20. Retrieved April 22, 2019, from https://www.buzzfeednews.com/article/craigsilverman/partisan-fb-pages-analysis.

Spivak, C. (2010). The fact-checking explosion: In a bitter political landscape marked by rampant allegations of questionable credibility, more and more news outlets are launching truth-squad operations. *American Journalism Review, 32*(4), 38–44.

Stencel, M. (2019). Number of fact-checking outlets surges to 188 in more than 60 countries. *Poynter*. Retrieved August 6, 2019, from https://www.poynter.org/fact-checking/2019/number-of-fact-checking-outlets-surges-to-188-in-more-than-60-countries/.

Sugars, S. (2019). From fake news to enemy of the people: An anatomy of Trump's tweets. *Committee to Protect Journalists*, January 30. Retrieved June 8, 2019, from https://cpj.org/blog/2019/01/trump-twitter-press-fake-news-enemy-people.php.

Szpitalak, M., & Polczyk, R. (2010). Warning against warnings: Alerted subjects may perform worse misinformation, involvement and warning as determinants of witness testimony. *Polish Psychological Bulletin, 41*(3), 105–112.

Szpitalak, M., & Polczyk, R. (2011). Can warning harm memory? The impact of warning on eyewitness testimony. *Problems of Forensic Sciences, 86*, 140–150.

Szpitalak, M., & Polczyk, R. (2012). When does warning help and when does it harm? The impact of warning on eyewitness testimony. *Roczniki Psychologiczne/Annals of Psychology, 15*(4), 51–72.

Tandoc, E. C., Jr., Lim, Z. W., & Ling, R. (2018). Defining "fake news" a typology of scholarly definitions. *Digital Journalism, 6*(2), 137–153.

Thorson, E. (2016). Belief echoes: The persistent effects of corrected misinformation. *Political Communication, 33*(3), 460–480.

Tucker, J., Guess, A., Barberá, P., Vaccari, C., Siegel, A., Sanovich, S., et al. (2018). Social media, political polarization, and political disinformation: A review of the scientific literature. Report. *William and Flora Hewlett Foundation*. Retrieved December 17, 2019, from https://eprints.lse.ac.uk/87402/1/Social-Media-Political-Polarization-and-Political-Disinformation-Literature-Review.pdf.

van der Linden, S., Leiserowitz, A., Rosenthal, S., & Maibach, E. (2017). Inoculating the public against misinformation about climate change. *Global Challenges, 1*(2), 1–7.

Walker, M., & Gotttfried, J. (2019). Republicans far more likely than democrats to say fact-checkers tend to favor one side. *Pew Research Center*, June 27. Retrieved July 9, 2019, from https://www.pewresearch.org/fact-tank/2019/06/27/republicans-far-more-likely-than-democrats-to-say-fact-checkers-tend-to-favor-one-side/.

Wardle, C. (2017). Fake news. It's complicated. *FirstDraft*, February 16. Retrieved April 22, 2019, from https://medium.com/1st-draft/fake-news-its-complicated-d0f773766c79.

Wardle, C., & Derakhshan, H. (2017). *Information disorder: Toward an interdisciplinary framework for research and policy making* (p. 9). DGI: Council of Europe report.

Weeks, B. E. (2015). Emotions, partisanship, and misperceptions: How anger and anxiety moderate the effect of partisan bias on susceptibility to political misinformation. *Journal of Communication, 65*(4), 699–719.

Wintersieck, A., Fridkin, K., & Kenney, P. (2018). The message matters: The influence of fact-checking on evaluations of political messages. *Journal of Political Marketing*. https://doi.org/10.1080/15377857.2018.1457591.

Wong, T. (2019). Singapore fake news law polices chats and online platforms. *BBC News*, May 9. Retrieved June 8, 2019, from https://www.bbc.com/news/world-asia-48196985.

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

Wyler, H., & Oswald, M. E. (2016). Why misinformation is reported: Evidence from a warning and a source-monitoring task. *Memory, 24*(10), 1419–1434.

Zaller, J. R. (1992). *The Nature and Origins of Mass Opinion*. Cambridge, UK: Cambridge University Press.

Zaragoza, M. S., Belli, Robert F., & Payment, K. E. (2006). Misinformation effects and the suggestibility of eyewitness memory. In Garry, M. and Hayne, H. (Eds.), *Do justice and let the sky fall: Elizabeth F. Loftus and her contributions to science, law, and academic freedom* (pp. 35–63). Hillsdale, NJ: Lawrence Erlbaum Associates.

Affiliations

Melanie Freeze[1] · Mary Baumgartner[2] · Peter Bruno[3] · Jacob R. Gunderson[4] · Joshua Olin[5] · Morgan Quinn Ross[6] · Justine Szafran[7]

Mary Baumgartner
baumgartner.mary5@gmail.com

Peter Bruno
peterbruno13@gmail.com

Jacob R. Gunderson
jacob.gunderson@unc.edu

Joshua Olin
josholin52@gmail.com

Morgan Quinn Ross
ross.1655@osu.edu

Justine Szafran
justineszafran@gmail.com

1 Department of Political Science, Carleton College, One North College Street, Northfield, MN 55057, USA

2 Minneapolis, USA

3 Washington, USA

4 Department of Political Science, University of North Carolina at Chapel Hill, 361 Hamilton Hall, CB 3265, Chapel Hill 27599-3265, North Carolina, USA

5 New York, USA

6 School of Communication, Ohio State University, 3016 Derby Hall, 154 N Oval Mall, Columbus 43210, Ohio, USA

7 Wilmette, USA